Book of Abstracts

Edited by Eetu Mäkelä, Mikko Tolonen and Jouni Tuominen

This book contains all abstracts presented at the DHN2018 conference. For proposals submitted as publication ready texts, see the citable proceedings at http://ceur-ws.org/Vol-2084/.

The following two pages contain the table of contents for this book. To locate the individual abstracts, use the bookmark functionality of your PDF reader.

Paper ID	Authors	Title
100	Snickars, Pelle	Breaking Bad (Terms of Service)? The DH-scholar as Villain
104	Huvila, Isto	The big challenge of data! Managing digital resources and infrastructures for digital humanities researchers
106	Porwoł, Monika	Synergy of contexts in the light of digital humanities: a pilot study
109	Hrafnkelsson, Örn	Ownership and geography of books in mid-nineteenth century Iceland
112	van Lange, Milan, Futselaar, Ralf	Dialects of Discord. Using word embeddings to analyze preferred vocabularies in a political debate: nuclear weapons in the Netherlands 1970-1990
113	Graham, Elyse	“Database Thinking and Deep Description: Designing a Digital Archive of the National Synchrotron Light Source (NSLS)”
116	Kekki, Saara	Network Analysis, Network Modeling, and Historical Big Data: The New Networks of Japanese Americans in World War II
118	Hämäläinen, Lasse	Names as a Part of Game Design
122	Fratiloiu, Raluca	The Stanley Rhetoric: A Procedural Analysis of VR Interactions in 3D Spatial Environments of Stanley Park, BC
124	Burrows, Toby	Cultural heritage collections as research data
125	Wijsman, Hanno, Hyvönen, Eero, Ransom, Lynn, Burrows, Toby	Big Data and the Afterlives of Medieval and Renaissance Manuscripts
126	Deeming, Karen Lisa	The New Face of Ethnography: Utilizing Cyberspace as an Alternative Study Site
133	Nielbo, Kristoffer Laigaard, Malm, Mats, Thomsen, Mads Rosendahl	Research in Nordic literary collections: What is possible and what is relevant?
136	Härkönen, Antti	Using ArcGIS Online and Story Maps to visualise spatial history: The case of Vyborg
137	Roine, Hanna-Riikka	The Future of Narrative Theory in the Digital Age?
139	Martin, Benjamin G.	Charting the ‘Culture’ of Cultural Treaties: Digital Humanities approaches to the history of international ideas
143	Naukkarinen, Ossi, Pacauskas, Darius	Finnish aesthetics in scientific databases
145	Norén, Fredrik, Mähler, Roger	The World According to the Popes: A Geographical Study of the Papal Documents, 2005–2017
146	Kivioja, Virpi, Elo, Kimmo	EXPLORING COUNTRY IMAGES IN SCHOOL BOOKS: A COMPARATIVE COMPUTATIONAL ANALYSIS OF GERMAN SCHOOL BOOKS IN THE 20TH AND THE 21ST CENTURY
148	Anderson, Rebecca	The Science of Sub-creation: Transmedial World Building in Fantasy-Based MMORPGs
150	Dufva, Tomi	Its your data, but my algorithms
151	Svedjedal, Johan, Stymne, Sara, Östman, Carin	Prose Rhythm in Narrative Fiction: the case of Karin Boye's Kallocain
152	Nauha, Tero	Digital archives and the learning processes of performance art
153	Pihlflyckt, Katarina	Approaching a digital scholarly edition through metadata
154	Stam, Per, Claes, Pieter, Veit, Elisa Johanna	Challenges in textual criticism and editorial transparency
156	Vaara, Ville, Vesanto, Aleksi, Tolonen, Mikko, Sippola, Reetta, Salmi, Hannu, Rantala, Heli, Ginter, Filip, Nivala, Asko, Mäkelä, Eetu, Marjanen, Jani, Lahti, Leo, Kanner, Antti	Metadata Analysis and Text Reuse Detection: Reassessing public discourse in Finland through newspapers and journals 1771–1917
158	Vaivade, Anita, Abele, Liga	Legal issues regarding tradition archives: the Latvian case study.
162	Olstad, Vemund, Olsson, Anders	Heritage Here, K-Lab and intra-agency collaboration in Norway
165	Treija, Rita, Daugavietis, Jānis	CAWI for DH
168	Camps, Martin	“Memes” as a Cultural Software in the Context of the (Fake) Wall between the US and Mexico
171	Navarra, carlo, Neset, Tina-Simone, Käyhkö, Janina, Asplund, Therese, Juhola, Sirkku	Serious gaming to support stakeholder participation and analysis in Nordic climate adaptation research
172	Wessman, Anna, Thomas, Suzie, Tuominen, Jouni, Koho, Mikko, Salmela, Ulla, Rohiola, Ville, Hyvönen, Eero, Ikkala, Esko	SuALT: Collaborative Research Infrastructure for Archaeological Finds and Public Engagement through Linked Open Data
173	Baunvig, Katrine Frøkjær, Nielbo, Kristoffer Laigaard	The Dostoyevskian Trope: State Incongruence in Danish Textual Cultural Heritage
175	Garnett, Vicky, Nugent Folan, Georgina, Edmond, Jennifer	When Open becomes Closed: Findings of the Knowledge Complexity (KPLEX) Project.
177	Edoff, Erik	A newspaper atlas: Named entity recognition and geographic horizons of 19th century Swedish newspapers
178	Tolonen, Mikko, Lahti, Leo, Mäkelä, Eetu, Vaara, Ville, Marjanen, Jani, Kanner, Antti, Hill, Mark J.	Spheres of “public” in eighteenth-century Britain
179	Torpus, Jan	Extending museum exhibits by embedded media content for an embodied interaction experience
180	Turtiainen, Riikka, Vaahensalo, Elina, Östman, Sari	Where are you going, research ethics in Digital Humanities?
181	Charbonneau, Olivier	Copyright exceptions or licensing : how can a library acquire a digital game?
182	Stubb, Maria Elisabeth	Shearing letters and art as digital cultural heritage, co-operation and basic research
184	Hella, Anni, Vesanto, Aleksi, Kaartinen, Marjo	Revisiting the authorship of Henry VIII’s Assertio septem sacramentorum through computational authorship attribution
187	Pálsson, Gísli, Roued, Henriette, Huggett, Jeremy, Paliou, Eleftheria, Traviglia, Arianna, Davidovic, Antonia, Laužikas, Rimvydas, Thomas, Suzie, Dallas, Costis, Huvila, Isto	ARKWORK: Archaeological practices and knowledge in the digital environment
188	Smith, Marcus, Källström, Magnus, Bianchi, Marco	"Everlasting Runes": A Research Platform and Linked Data Service for Runic Research
189	Laakkonen, Simo	Digital humanities and environmental reporting in television during the Cold War Methodological issues of exploring materials of the Estonian, Finnish, Swedish, Danish, and British broadcasting companies
190	Onikki-Rantajääskö, Tiina, Enqvist, Johanna	The Bank of Finnish Terminology in Arts and Sciences – a new form of academic collaboration and publishing

193	Scherrer, Yves, Samardžić, Tanja	ArchiMob: A multidialectal corpus of Swiss German oral history interviews
194	Mähler, Roger, Steinvall, Anders, Svensson, Jon, Lindvall-Östling, Mattias, Deutschmann, Mats	“See me! Not my gender, race, or social class”: Combating Stereotyping and prejudice mixing digitally manipulated experience with classroom debriefing.
195	Välimäki, Reima	Refutatio errorum – authorship attribution on a late-medieval antiheretical treatise
197	Revuelta-Eugercios, Barbara, Tovgaard-Olsen, Katrine, Clausen, Nanna Floor	From crowdsourcing cultural heritage to citizen science: how the Danish National Archives 25-year old transcription project is meeting digital historians
198	Turunen, Risto	Sculpting Time: Temporality in the Language of Finnish Socialism, 1895–1917
204	Ruckenstein, Minna	Broken data and repair work
205	Matres, Inés	A long way? Introducing digitized historic newspapers in school, a case study from Finland
211	Watrall, Ethan	Towards an Approach to Building Mobile Digital Experiences For University Campus Heritage & Archaeology
212	Kelomees, Raivo	Art of the Digital Natives and Predecessors of Post-Internet Art
213	Rees, Ellen	A Computational Assessment of Norwegian Literary “National Romanticism”
217	Latva, Otto, Nivala, Asko, Oiva, Mila, Salmi, Hannu	Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914
219	Arjava, Heini	Prosodic clashes between music and language – challenges of corpus-use and openness in the study of song texts
220	Paju, Petri, Oiva, Mila, Fridlund, Mats	Facilitating Digital History in Finland: What can we learn from the past?
222	Vesanto, Aleksi, Vaara, Ville, Tolonen, Mikko	Text Reuse and Eighteenth-Century Histories of England
223	Mähler, Roger, Jarlbrink, Johan	Embedded words in the historiography of technology and industry, 1931–2016
231	Mäkelä, Eetu, Kanner, Antti, Marjanen, Jani, Hengchen, Simon	Comparing Topic Model Stability Between Finnish, Swedish and French
235	Tennøe, Arthur, Johnsen, Lars Bagøien	Making a bibliography using metadata
236	Liljestrand, Niklas	Designing a Generic Platform for Digital Edition Publishing
237	Korkiakangas, Timo	Network visualization for historical corpus linguistics: externally-defined variables as node attributes
240	Karlsen, Heidi	Interdisciplinary advancement through the unexpected: Mapping gender discourses in Norway (1840-1913) with Bokhylla
241	Lehtiniemi, Tuukka, Ruckenstein, Minna	Shaping data futures: Towards non-data-centric data activism
242	Ånäs, Susanna	Wikidocumentaries
243	Niskanen, Samu Kristian, Leinonen, Lauri Iisakki	Medieval Publishing from c. 1000 to 1500
244	Glebova, Daria	Using rolling.classify on the Sagas of Icelanders: Collaborative Authorship in Bjarnar saga Hítdælakappa
245	Kanner, Antti	Two cases of meaning change in Finnish newspapers, 1820-1910
246	Kallio, Maria	Handwritten Text Recognition and 19th Century Court Records
247	Muischnek, Kadri, Lust, Kersti, Lindström, Liina, Jaanimäe, Gerth, Pilvik, Maarja-Liisa	Creating a corpus of communal court minute books: a challenge for digital humanities
250	Vats, Ekta, Hast, Anders, Mårtensson, Lasse	Extracting script features from a large corpus of handwritten documents
251	Leinonen, Lauri, Eskola, Seppo	Diplomatarium Fennicum and the digital research infrastructures for medieval studies
252	Holloway-Attaway, Lissa	Critical Play, Hybrid Design and the Performance of Cultural Heritage Game/Stories
255	La Mela, Matti	Digitised newspapers and the geography of the nineteenth-century “lingonberry rush” in Finland
258	Roivainen, Hege	Identifying poetry based on library catalogue metadata
264	Rašmane, Anita, Bojārs, Uldis	Semantic Annotation of Cultural Heritage Content
265	Fewster, Derek	Layers of History in Digital Games
266	Laak, Marin, Viires, Piret	Digital Humanities Meets Literary Studies: the Challenges for Estonian Scholarship
274	Dellner, Jennifer J	The plague transformed: City of Hunger as mutation of narrative and form
276	Niku, Maria, Keravuori, Kirsi	Elias Lönnrot Letters Online
277	Korpijärvi, Anna-Leena, Airaksinen, Tiina H.	KuKa Digi -project
278	Gritsenko, Daria, Isoaho, Karoliina	Topic modelling and qualitative textual analysis
279	Kokko, Heikki	Local Letters to Newspapers - Digital History Project
280	Poder, Søren	Lessons Learned from Historical Pandemics. Using crowdsourcing 2.0 and Citizen Science to map the Spanish Flus spatial and social network.
281	Kivelä, Mikko, Asikainen, Aili, Kaski, Kimmo, Iñiguez, Gerardo	Triadic closure amplifies homophily in social networks
282	Fredriksson, Anna Cecilia	Dissertations from Uppsala University 1602-1855 at the internet
283	Mäkelä, Eetu, Säily, Tanja, Hämäläinen, Mika	Normalizing Early English Letters for Neologism Retrieval
284	Eide, Stian Rødven, Borin, Lars, Tahmasebi, Nina, Rouces, Jacobo	Analysing Swedish Parliamentary Voting Data
285	Hämäläinen, Mika, Soisalon-Soininen, Eliel	Automated Cognate Discovery in the Context of Low-Resource Sami Languages

Digital Humanities in the Nordic Countries, Helsinki, Finland, 7–9 March 2018.

Breaking Bad (Terms of Service)? The DH-scholar as Villain

Professor Pelle Snickars
Department of Culture and Media Studies / Humlab
Umeå university
pelle.snickars@umu.se

For a number of years I have been heading a major research project on Spotify (funded by the Swedish Research Council). Entitled, "Streaming Heritage: Following Files in Digital Music Distribution", the project has involved system developers Fredrik Palm, Roger Mähler, Andreas Marklund and Johan von Boer (at Humlab, Umeå University), as well as researchers Maria Eriksson, Anna Johansson (at Umeå University), Rasmus Fleischer (at Stockholm and Umeå University), and Patrick Vonderau (at Stockholm University). A guiding question for the project has been how people's practices and approaches toward cultural forms such as songs, books, or films—practices including the production, expression, and exchange of those cultural forms—are currently transformed under the shift from commodity ownership to digitized and commodified experiences (with Spotify as our prime example).

The project has taken a software studies and digital humanities approach towards streaming media. It has repeatedly engaged in reverse engineering Spotify's algorithms, aggregation procedures, and valuation strategies—and has hence repeatedly been non-compliant with Spotify's Terms of Service (ToS). Importantly, however, the project has been explicit and open about its methods and critical approach from the very start (in 2014). The project's research forms part of a tradition, well established within the social sciences and the humanities, which studies media actors critically. Studying industries critically means that our research is not *applied* research, that is, not conducted on any company's behalf against payment—but in the argumentative form and based on the standards of problem-oriented humanist inquiry. Different forms of public activism have, for example, been part of the project's 'interventionist' research design. In blog posts, newspaper articles, public service interviews—in both television and radio—the project's interventionist strategy has been outlined as based on digital methods, including the use of bots and the dissemination of self-produced sounds via Spotify.

The project in many ways followed calls to use the same tools that organize online information, due to Spotify's reluctance to share data. Our encounters with Spotify began with a brief conversation with (then) Head of Marketing, Sophia Bendz, in Stockholm in the fall of 2012. Asked if Spotify would not qualify as a regular media company—given its declared business interest of providing content to audiences, while selling those audiences to advertisers—Bendz rushed to praise Spotify's achievements as a tech company. A follow-up meeting with a company executive in Spotify's headquarters (before having received the grant funding), led to mutual expressions of respect, but little more. At this point, Spotify had already been made aware of our research interest to work with the company's data. Over the years, we individually met with engineers and marketers, data geeks and academics related to the company. But conversations often fell apart as soon as Spotify's 'tech identity' was questioned.

This inspired a research approach that would go beyond interviews, direct observation, and other standard methods of media studies research. Not that we researchers avoided talking to Spotify employees—we met a few of them over the years—but as with any other media company striving for a monopoly position, listening to company spin and "industrial self-theorizing" felt not enough. Complementing such 'frontend' inquiries with experimental 'backend' studies of digital media infrastructure, metadata generation, and aggregation practices, the project aimed to initiate public debate about the often subtly changing standards, values, and politics of cultural dissemination online.

Spotify once offered a technical solution for music distribution, yet the aggressive discursive framing of Spotify's operation as being primarily technological has tended to obscure its long-term entrepreneurial, financial, and culture-changing strategies. Since 2015, for instance, Spotify has implemented a plan—and the technology—to generate data based on its music streaming that allow to study human behavior at scale. The company acts, in other words, not only as a music provider, but also as a private data broker.

In order to study such data flows, our project's key methodological suggestion has been to 'follow files' rather than those making, using or collecting them. Through the notion of 'following files' our project developed a habitual way of working with digital methods at Humlab (Umeå University), to understand both Spotify's general streaming architecture, as well as particularities of the service interface, its music recommendation engine, or radio functionality. We deployed different types of digital methods—from computational tracking of all data transmissions that occur every time one presses 'play', to working with bots as virtual listeners that logs their 'behaviours', or text mining approaches of scraped Spotify job advertisements. In addition, the way in which we have studied music aggregation from the inside—by launching our own record label—has been yet another way of looking closer at some of the entries (or 'holes') in the 'Spotify black box', not the least by getting access to different monitoring services at aggregators, or concretely learning what kind of metadata categories that are available (and obligatory) when music is being bundled into packages of differentiated data. A first thing to note, for example, when going *under the hood* is that the Spotify infrastructure hardly appears as an uniform platform. Rather it is downright traversed by data flows, file transfers and information retrieval in all kinds of directions—be they metadata traffic identifying music, aggregation of audio content, playout of streaming audio formats (in different quality ratings), programmatic advertising (modelled on finance's stock exchanges) or interactions with other services (notably social media platforms). Spearheading the new data economy of the 21st century, Spotify in many ways resembles a sprawling network of interaction that includes musicians and listeners alongside other actors and interests that have little to do with cultural commodities or media markets in a traditional sense.

During the summer of 2017 I received an email from a Spotify legal counsel who was "concerned about information it received regarding methods used by the responsible group of researchers in this project. This information suggests that the research group systematically violated Spotify's Terms of Use by attempting to artificially increase plays, among others, and to manipulate Spotify's services with the help of scripts or other automated processes." I responded politely—telling him what we had done, and that we would gladly share our research results with Spotify. I got no answer. A few weeks later, however, I received a letter from the senior legal advisor at my university. Spotify had apparently contacted the Swedish Research Council with the claim that our research was questionable in a way that would demand "resolute action", and the possible termination of financial and institutional support. Our research group were asked to formally describe and rebut the claims—we wrote:

"Let us begin by firmly stating that neither the aim, methods nor results of our project were in any way designed or used to cause harm to Spotify or any of its users—or to benefit commercially from non-authorized access to the service's proprietary data. Our results do not reveal any detailed information about Spotify's proprietary algorithms or software, or disclose information that might be harmful if it ends up in the hands of Spotify's competitors. Our scientific research and its scholarly findings are thus not a competitive threat to Spotify as a company. We have never violated the integrity of any Spotify user, or collected any personal data related to Spotify users, or illegally shared copyrighted content via Spotify. We have respected Spotify's—and any other company's—wishes to protect the integrity of their service and brand. […] Spotify mixes three very different standards in making its claim: *methodical*, *ethical*, and *legal* standards. Spotify is concerned about a possible violation of its Terms of Service (ToS). The framework of *user agreements* (or ToS) is all Spotify has to make this claim. While Spotify's insistence on our adherence to its user agreements is well-founded, it is also off the point as we have ended all activities that could be understood as being in violation of these agreements"

DH-research is embedded in 'the digital'—and so are its methods, from scraping web content to the use of bots as research informants. Within scholarly communities centered on the study of the web or social media there is a rising awareness of the ways in which digital methods might be non-compliant with commercial Terms of Service (ToS)—a discussion which has not yet filtered out and been taken serious within the digital humanities. However, DH-researchers will in years to come increasingly have to ask themselves if their scholarly methods need to abide by ToS—or not. As Amy Bruckman has stated, it might have profound scholarly consequences: "Some researchers choose not to do a particular piece of work because they believe they can't violate ToS, and then another researcher goes and does that same study and gets it published with no objections from reviewers."

In September 2017, the Swedish Research Council as well as Umeå University decided to close the case; Spotify's emailed request were thus turned down. My paper will recount our legal dealings with Spotify—including a discussion of the digital methods used in our project—but also more generally reflect around the ethical implications of collecting data in novel ways. ToS are contracts—not the law. Still there is a dire need for ethical justifications and scholarly discussions why the importance of academic research justifies breaking ToS.

Keywords: digital methods, violation of Terms of Service, research ethics, scholarly reviews

The big challenge of data! Managing digital resources and infrastructures for digital humanities researchers

Isto Huvila

Department of ALM, Uppsala University
firstname.lastname@abm.uu.se

Abstract

Digital humanities research is dependent on the development and seizing of appropriate digital methods and technologies, collection and digitisation of data, and development of relevant and practicable research questions. In the long run, the potential of the field to sustain as a significant social intellectual movement (or in Kuhnian terms, paradigm) is, however, conditional to the sustainability of the scholarly practices in the field. Digital humanities research has already moved from early methodological experiments to the systematic development of research infrastructures. These efforts are based both on the explicit needs to develop new resources for digital humanities research and on the strategic initiatives of the keepers of relevant existing collections and datasets to open up their holdings for users. Harmonisation and interoperability of the evolving infrastructures are in different stages of developments both nationally and internationally but in spite of the large number of practical difficulties, the various national, European (e.g. DARIAH, CLARIN and ARIADNE) and international initiatives are making progress in this respect. The sustainability of digital infrastructures is another issue that has been scrutinised and addressed both in theory and practice under the auspices of national data archives, specialist organisations like the British Digital Curation Centre and international discussions, for instance, within the iPRES conference community. However, an aspect of the management of the infrastructures that has received relatively little attention so far, is management for use. We are lacking a comprehensive understanding of how the emerging digital data and infrastructures are used, could be used and consequently, how the emanating resources should be managed to be useful for digital humanities research not only in the context within which they were developed but also for other researchers and many cases users outside of the academia.

This paper discusses the processes and competences for the management of digital humanities resources and infrastructures for maximising their current and future usefulness for the purposes of research. On the basis of empirical work on archaeological research data in the context of the Swedish Archaeological Information in the Digital Society (ARKDIS) research project (Huvila, 2014) and a comparative study with selected digital infrastructures in other branches of humanities research, a theoretical model of use-oriented management of research

data with central processes and competences is presented. The suggested approach complements existing digital curation and management models by opening up for the consideration of the user side processes of digital humanities data resources and their implications for the functioning, development and management of appropriate research infrastructures. Theoretically the approach draws from the records continuum theory (as formulated by Upward and colleagues (e.g. Upward, 1996, 1997, 2000; McKemmish, 2001)) and Pickering's notion of the mangle of practice (Pickering, 1995) developed in the context of the social studies of science. The model demonstrates the significance of being sensitive to explicit wants and needs of the researchers (users) but also the implicit, often tacit requirements that emerge from their practical research work. Simultaneously, the findings emphasise the need of a meta-competence to manage the data and provide appropriate services for its users.

Bibliography

Huvila, I. (Ed.) (2014). *Perspectives to Archaeological Information in the Digital Society*. Uppsala: Department of ALM, Uppsala University.
URL http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-240334

McKemmish, S. (2001). Placing Records Continuum Theory and Practice. *Archival Science*, *1*(4), 333–359.
URL http://dx.doi.org/10.1023/A:1016024413538

Pickering, A. (1995). *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

Upward, F. (1996). Structuring the Records Continuum Part One: Post-custodial Principles and Properties. *Archives and Manuscripts*, *24*(2), 268–285.

Upward, F. (1997). Structuring the Records Continuum, Part Two: Structuration Theory and Recordkeeping. *Archives and Manuscripts*, *25*(1), 10–35.

Upward, F. (2000). Modelling the continuum as paradigm shift in recordkeeping and archiving processes, and beyond–a personal reflection. *Records Management Journal*, *10*(3), 115–139.

Synergy of contexts in the light of digital humanities: a pilot study

Monika Porwoł

Abstract. The second generation humanities pertains to a synergy of the humanistic thought and knowledge with the digital tools that deal with ʻbig-dataʻ analysis (i.e. processing, visualization, presentation and popularization of research results). This short paper describes a pilot study pertaining to the linguistic analysis of meaning with regard to the word ladder[EN]/drabina[PL] taking into account views of digital humanities. Therefore, WordnetLoom mapping is introduced as one of the existing research tools proposed by CLARIN ERIC research and technology infrastructure. It is a lexico-semantic network, in which the lexeme ladder/drabina is defined implicitly in reference to other words and presented in a ʻconcept mapʻ via WordNetLoom Viewer (an application enabling display of plWordNet entries). The semantic network of the word ladder/drabina represents semantic relations, i.e. synonymy {partial, (inter)register, (inter)pragmatic}, hyponymy, holonymy, meronymy, hyperonymy. A path to the highest hyperonym is constructed in the following order: an entity, physical entity, object, artifact, stairway. Moreover, the explicated material comprises retrospective remarks and interpretations provided by 74 respondents, who took part in a survey. A detailed classification of multiple word's meanings is presented in a tabular way (showing the number of contexts, in which participants accentuate the word ladder/drabina) along with some comments and opinions. Undoubtedly, the results suggest that apart from the general domain of the word offered for consideration, most of its senses can usually be attributed to linguistic recognitions. A great part of survey participants was able to unify the meanings of the word ladder/drabina by matching a direct sense out of the provided collection of word meanings in the adequate context. Language users can easily determine the denotative meanings of the word, as well as exemplify that lexeme in connotative sentences by illustrating their point through metaphors or idiomatic expressions. However, in several instances the answers were marked by doubt due to a lack of knowledge, uncertainty or carelessness in answering the questions. In other words, the majority of subjects took notice of the general understanding of the word ladder/drabina in an ordinary domain and then proceeded to its interrelated senses (depending on the personal wisdom and understanding of concepts). The linguistic background, however, was indispensable in determining the word's annotations.

Keywords: digital humanities, CLARIN ERIC, plWordNet, WordnetLoom Viewer, linguistic analysis, pilot study.

Ownership and geography of books in mid-nineteenth century Iceland

Abstract

In October 1865, the national librarian, then the only employee of the National Library of Iceland (est. 1818) got permission from the bishop in Iceland to send a written request to all provosts around the country to do a detailed survey in the parishes of ownership of old Icelandic books printed before 1816. The title page of every book in all farms should be written down in full detail, the number of printed pages recorded, and the place of publication *etc*. The aim of this project was to compile data for a detailed national bibliography with a list of Icelandic authors, and to build a good collection of books in the library. Some of the written reports have survived and are now in the library archive. In this paper, I will talk about these unused sources of books ownership and book history. Most importantly, how we are using them with other data to display how ownership of books in the mid-nineteenth century for example varied indifferent parts of the country. By using several authority files that are in TEI P5 XML a detailed map of ownership can be drawn. The authority files are firstly, a detailed historical bibliography of Icelandic books from 1534 to 1844 and secondly a list of all farms in Iceland with GPS coordinates. In my talk, I will also elaborate on how this project about ownership of books and the geography of books can be developed further and how the data can be used by others. One aspect of my short talk is the cooperation between librarians and IT professionals and how unrelated sources can be linked together to bring out new knowledge and interpret history.

Projects website: https://bokaskra.landsbokasafn.is/geography

ABSTRACT Helsinki DHN 2018

Ralf Futselaar

Milan van Lange

Dialects of Discord

Using Word Embedding Models to analyse preferred vocabularies in political debate through time: the nuclear weapons controversy in the Netherlands, 1970-1990.

Introduction

We analyse the debate about the placement of nuclear-enabled cruise missiles in the Netherlands during the 1970s and 1980s. The NATO "double-track decision" of 1979 envisioned the placement of these weapons in the Netherlands, to which the Dutch government eventually agreed in 1985. In the early 1980s, the controversy regarding placement or non-placement of these missiles led to the greatest popular protests in Dutch history and to a long and often bitter political controversy. After 1985, due to declining tensions between the Soviet Block and NATO, the cruise missiles were never stationed in the Netherlands. Much older nuclear warheads, in the country since the early 1960s, remain there until today.[1]

We are using Word Embedding Models (WEMs) to analyse this acrimonious debate in the proceedings of the Dutch lower and upper house of Parliament. The official political positions either for or against deployment, as expressed in party manifestos and voting behaviour inside parliament, were stable throughout this period. We demonstrate that in spite of this apparent stability, the vocabularies used by representatives of different political parties changed significantly through time.

Data

In this investigation we used a dataset of parliamentary records, known in Dutch as the *Handelingen der Staten-Generaal*. The data contains the verbatim minutes of both houses of parliament for the period 1814-1995. The dataset was first digitized by the Royal Library of the Netherlands and dramatically improved in the *Political Mashup*-project that ran from 2012 to 2016. The dataset consists of a large collection of XML files containing the complete minutes of all the meetings of the lower and upper chambers of parliament, separated by date, topic, political affiliation, etc. The corpus is not only sufficiently comprehensive for the creation of WEMs (an average of 1266 documents/topics per year for the period under scrutiny), but also complete. This makes it an excellent corpus for various forms of automated text analysis.

[1] Remco van Diepen, *Hollanditis? Nederland en het kernwapendebat 1977-1987* (Amsterdam: Boom, 2004); Jan Hoffenaar, *Confrontatie en ontspanning. Maatschappij en krijgsmacht in de Koude Oorlog 1966-1989* (Den Haag: SDU, 2004).

Method

We have used the *Handelingen* and Google's word2vec algorithm to train our WEMs. We started with the Dutch word for "nuclear weapon" ("kernwapen") and identified its nearest neighbours (words close in vector space). We manually annotated all synonyms and near- synonyms of "nuclear weapon" used in the proceedings of both houses of parliament during the period under scrutiny. By using the mean of all the vectors of these words, we have created a combined vector representing the concept, rather than the term, of "nuclear weapon" in vector space. Based on this combined vector we have identified nearest neighbours of words used to refer to nuclear weapons inside the WEMs. These terms have been manually classified, insofar relevant, into terms associated with a pro-proliferation or anti-proliferation viewpoint, for example "defense" and "disarmament" respectively. We used these viewpoint-associated words to create a pro-proliferation combined vector, and an anti-proliferation one.

From the perspective of historical research, WEMs have a fundamental weakness. Word vectors can only be compared with other vectors within the same spatial model. The same word in a different model may (and will) have an entirely different vector inside another model. Since change through time is the core of virtually all historical research (including this investigation), this presents us with a major problem; how can we compare outcomes for different periods in time? For this investigation, we have created a number of shifting windows, overlapping corpora which each contain the data from a specific time period of five years. We trained WEMs on each corpus. As explained, vectors from each of these separately trained models cannot as such be compared. To overcome this problem, we have used 250 words closely associated with the combined vector of "nuclear weapon" and calculated the cosine similarity of each of these 250 word vectors to the combined vectors representing the two different viewpoints (pro-proliferation and anti-proliferation). We repeated this for each of the models/windows. With these cosine-similarity-scores, we have created a measure to compare word vectors' closeness outside a single model.

This allows us to compare political discourses in discussions on nuclear weapons in relation to the two possible viewpoints in this bipolar debate. It is worthy to note that we have chosen a two-dimensional implementation in this case, but that this is not theoretically necessary. Using two viewpoints, each represented by a combined vector, does allow us to project the closeness of the top 250 nuclear-weapon-related words in each of the periods in a two-dimensional space in which one viewpoint serves as an y axis, the other as a x axis. Thus, we have arrived at a visual representation that allows for a comparison of closeness in WEMs for more than one corpus and hence for a comparison through time. Needless to say, it would be necessary, when using this approach, to use relatively similar corpora. For historical research into relatively short periods of parliamentary history, this is not particularly problematic.

Results

Obviously, representatives of all Dutch political parties used words from both categories in parliamentary debates. It is almost impossible, after all, to engage in a discussion without ever using the terms used by your adversaries. At any given time, however, we demonstrate that different political parties could be shown to have clear preferences in terms of vocabulary. In the "discursive space" created by the binary opposition between pro- and contra-proliferation words, political parties can be shown to have had specific and distinct ways of discussing nuclear weapons. We identify and distinguish the specific vocabularies used by the Communist, Social Democrat, Christian Democrat and Liberal parties.

Using this framework, we have analysed the changing vocabularies of different political parties. This allows us to show that, while stated policy positions and voting behaviour remained unchanged, the language used to discuss nuclear weapons shifted strongly towards anti-proliferation terminology. We have also been able to show that this change happened at different times for different political parties. We speculate that these changes in the preferred "dialect" to discuss nuclear weapons resulted from perceived changes of opinion among the target electorates of different parties, as well as the changing geopolitical climate of the mid-to-late 1980s, where nuclear non-proliferation became a more widely shared policy objective.

Conclusion

The uptake of WEMs in industries that use text mining as an integral part of their business model (Google, Facebook, etc.) has been rapid and almost comprehensive. Their sudden popularity owes much to their proven effectiveness. In historical research, by contrast, they have thus far not been used very much. This is doubtlessly due to a degree, to lack of expertise and ingrained conservatism within the historical profession. A much greater hindrance, thus far, was the lack of usability of WEMs for diachronic analysis. We believe that we have proposed a methodology to use WEMs in a meaningful way for historical enquiry.

In the analysis of political speech, the resulting findings have a separate relevance. Political scientists often use vocabularies to predict or estimate political viewpoints. In the case at hand, we have demonstrated that when identified political viewpoints remain unchanged, politicians do adapt their vocabularies to appeal to the perceived views of their voters and potential voters. In this way, WEMs can help to analyse the ways in which political ideas are sold, successfully or unsuccessfully, to electorates.

1: Introduction

Over the past few years, research carried out at large-scale materials science facilities in the United States and elsewhere has undergone a phase transition that has affected its character and culture. In this talk, I'd like to briefly (1) describe this phase transition, (2) review the practical challenges it poses for historians, (3) review some potential digital tools that might respond to these challenges, and then (4) suggest some theoretical challenges posed by the emerging field of "database history."

2: The New Big Science

What we might call the Old Big Science was born, starting after the Second World War, with the creation of national laboratories. The primary dynamic of the Old Big Science was that the scale of its premier high-energy physics projects—instruments, collaborations, and the time-scale that projects followed—increased rapidly. In the New Big Science, which started in the 1980's and entailed, in part, a shift to materials science, instruments and collaborations do not get bigger and bigger; instead, the research ecosystem grows more complicated. That ecosystem involves more and more fields making use of national laboratories, a wider variety of instruments, more connections between seemingly disparate research programs, and a faster turnover of programs.

Research in the New Big Science differs from research in the Old Big Science in the *scope and complexity of its research networks*, which can simultaneously involve several national labs, several universities, and several industries. The *formation of knowledge* in this research culture is also different. Instead of research projects addressing single puzzles, materials scientists often work to pull together a mosaic of properties whose focus may change rapidly with the properties of the materials being produced.

3: Challenges for the Historian of the New Big Science

That description brings me to the focus of this paper: the challenges of telling the story of such a research ecosystem, and the opportunities that digital tools and methods offer for meeting such challenges.

Consider the traditional methods for investigating the research ecosystem at a science facility. The traditional tools that a historian can bring to such a research object are diverse, but narrow; confined to the length of a book or article, he must choose one or a few currents in the full ecosystem on which to focus. For example, he might choose to track a machine's *operational history*, its *administrative history*, or its *functional history*. He could focus on the publications associated with each port or research program; the personnel associated with each port or research program; the instruments associated with each port or research program, and when they were built, rebuilt, or replaced; the funding, in the form of prizes, grants, and other funding sources, associated with work associated with each port or research program; the industries associated with each port or research program; the historical map of research programs

at the NSLS, how long they lasted, and how they were funded; or the applications of research findings.

It is our argument, however, that narratives and lists of this kind are of little value unless they are connected with each other. So essential is interconnectedness to the New Big Science that traditional historical methods, which can follow only a few threads at a time, cannot adequately account for the functioning of this research ecosystem nor explain its achievements. This mode of practicing science can only be fully understood if connectivity and interdependence are placed in the foreground of the historian's attention. Because of the singular dependence of any one factor on an entire ecologic environment, even very basic and traditional questions about the new era of materials science are difficult to answer: for example, what biological science is happening at materials science facilities, and how has this science evolved? More detailed questions are still harder: what are the connections between specific kinds of biological research and other fields like physics, chemistry, and engineering? How have these connections evolved over time, and how are they related to the evolution of broader social concerns about health, epidemics, and so forth? These are interesting and important questions—not just to science historians, but also to scientists, administrators, funding agencies, and policymakers.

As it happens, we, too, live at a moment characterized by transformation in our forms of knowledge production, defined, in part, by new tools that facilitate the rethinking of research objects in terms of connections and relationships. In particular, these tools include the relational database and a growing humanistic concern with database use, an orientation that new media scholars have begun to address as *database thinking*.

4: Database Thinking and Historical Thinking

The best way to introduce the specific challenges of database thinking is to summarize the goals of our current project, which takes as its basis the National Synchrotron Light Source (NSLS). The NSLS was a historic site for groundbreaking research in materials science from 1982-2014; during that time, it was one of the world's most widely used scientific facilities. The NSLS is anchored in Brookhaven National Laboratory, and is managed on behalf of the US Department of Energy by Brookhaven Science Associates, of which our home institution of Stony Brook University is a partner. The machine spins electrons in a circle so that they give off light and then reaps the light through several dozen ports. At each port, instruments shape the light so that it can be put to a range of tasks, such as studying materials via electromagnetic radiation.

In terms of research, the NSLS is one of the most productive instruments ever built. The vast number of experiments that took place at the NSLS, and the vast amount of data that it produced during a history that encompassed many changing disciplines, make it nearly impossible to gain a comprehensive global view of the knowledge production that took place at this facility. Traditional historical methods and linear narratives fail to capture precisely the elements that demand new exploration. We are therefore collaborating to develop a new kind of digital tool to capture the history of this research. This project will comprise a digital archive to obtain a history of the NSLS, the main component of which will be a relational database that

allows for a wide range of forms of query and analysis. It's a traditional digital humanities project in the sense that the project aims to demonstrate that new kinds of digital tools can help with challenges of information overload by storing, integrating, and imaging the large amounts of information needed to answer the questions we may ask of a complex ecologic environment. The aim of the database will not be to present a block of answers to our own research questions, but rather to enable other historians to easily explore questions and historical narratives of their own. As we move forward, we intend to guide our progress with reference to models like Sein et al.'s Action Design Research, which offers a model for addressing design problems and theoretical problems as intertwined strands.[1]

However, these tools also pose a new set of problems—historiographic problems that concern the relationship between the domain expertise of the historian and the design of a historical database. As the creation of historical databases becomes an increasingly common practice, we have a growing need to articulate principles for the design of digital tools that will support the kinds of analysis and discovery that we have traditionally valued as historians. For example, while one benefit of using databases is supposed to be freedom from the limitations of received ideas—a notion that *Wired* magazine memorably publicized by heralding big data as "the end of theory"—scholars such as Ted Underwood have reminded us that our digital research tools rely, however invisibly, on a range of preexisting narratives and assumptions.[2] Moreover, as research objects in the digital humanities become more complex, we have a growing need to articulate principles for the design of digital infrastructures to support the management of new volumes and configurations of data relevant to humanists. Here, as with thick description, the deliberate negotiation of frameworks persists, albeit at layers sometimes invisible to the user. Call it deep *description*.

5: The Challenges of Database History

So for the closing three minutes, I would like to focus on a specific problem associated with the question of how to write a history of the New Big Science taking place at the NSLS. Specifically, we wish to focus on the question of "database history" as a genre. As Katherine Hayles discusses in her 2012 book, *How We Think: Digital Media and Contemporary Technogenesis*, as historians rely more and more on databases for the management and discovery of information, they must reckon in their methodology with the relationship between the architectures of narrative and database.[3] As we have just outlined, the problem in historiography that the new archive is designed to address lies in the narrow and linear structure of most historical narratives. But databases carry historiographic problems of their own. While the structure of a database evades traditional narrative elements such as sequence and fixed order, a database designed by and for historians may still capitulate to "narrative" in the sense of the embedded logics of the discipline. Does the design of a database, whether in the configuration of the user interface, the selection of its contents, or the indexes that manage its workings, limit the

1 Maung K. Sein et al., "Action Design Research," Management Information Systems Quarterly 35, 1 (2001).

2 Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," Wired Magazine (23 June 2008). Accessed online, 28 May 2016: http://www.wired.com/2008/06/pb-theory/. Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," Representations 127, 1 (Summer 2014): 64-72.

3 N. Katherine Hayles, *How We Think: Digital Media and Contemporary Technogenesis* (Chicago: The University of Chicago Press, 2012). Hayles proposes that we think of narrative and database as "symbionts," not opposites.

stories that emerge from its contents in a way that would be problematic in the terms that Ted Underwood has described in his criticism of algorithmic research: the limitations are unseen to the people performing the search, who think that the search process has no limitations?

A related question is: would the involvement of human agents in managing this database limit or predetermine the uses that will be made of its contents? This question regards the database as an analog not to a book, but to a journal like *Annales*, *Representations*, or *Année Sociologique*, which not only curated scholarship, but also famously nurtured distinctive schools of practice. Because the field of history, and the schools that comprise it, are under the control of a professional community of historians, the information structure that grounds a historical database will reflect a necessarily partial (and perhaps even political) set of interests. In other words: can there be any kind of database other than a politicized database?

Early answers to these kinds of questions suggest: probably not. But since this feature is part of what makes a historical database, as an artifact, evaluable as a contribution to the discipline, we might start looking for ways to embrace and foreground this inevitability as an opportunity. If database thinking entails thinking in terms of classifications and relationships, then the integration of database thinking with historiographic thinking represents, not an end of theory, nor (in a term that was famously applied prematurely to our postmodern technological society) an end of history,[4] but rather an extension of the historiographic tradition into the materially grounded, practice-based domain that Matt Ratto has called "critical making."[5]

4 On this subject, see Alan Liu, *Local Transcendence: Essays on Postmodern Historicism and the Database* (Chicago: University of Chicago Press, 2008).

5 Matt Ratto, "Critical Making: Conceptual and Material Studies in Technology and Social Life" (2011).

Network Analysis, Network Modeling, and Historical Big Data: The New Networks of Japanese Americans in World War II

In my ongoing doctoral research, I am employing network analysis and modeling to study the Japanese American incarceration in World War II (internment). Incarceration and the government-led dispersal of Japanese Americans disrupted the lives of some 110,000 people, including over 70,000 US citizens of Japanese ancestry, for the duration of the war and beyond. Many lost their former homes and enterprises and had to start their lives over after the war. Incarceration also had a very concrete impact on the communities: about 50% of those interned did not return to their old homes.

My research data consists of three different types of datasets from different sources:

1) The Japanese American Internee Data File, 1942-1946, freely available as a .PU file with accompanying code book at www.archives.gov. This dataset includes the background information of 109,000 incarcerated individuals, including their occupation, education, religion, etc. Individuals are grouped by family, making it possible to construct family groups. I have partly converted the coded data into an Excel spreadsheet using Excel formulae, while parts have been converted by a colleague through Python.
2) Final Accountability Roster for the Heart Mountain Relocation Center, 1945, freely available as PDF through www.heartmountain.org or www.ancestry.com, includes the end-of-incarceration information (address in camp, destination, type of leave) of 14,000 residents of the Heart Mountain camp in Wyoming.
3) Data extracted from the *Heart Mountain Sentinel*, incarceration camp newspaper that depicted various aspects of camp life, such as politics and social life (available through www.densho.org). This dataset forms the base for analyzing the networks, and is complemented by archival documents from various sources.

My study consists of five "subnetworks" (family, administrative-political, employment, social, and geospatial), which will finally be integrated into one full network as comprehensive as possible. In addition, I'm studying several emerging subgroups of people, especially women, individuals segregated for perceived disloyalty, and the draft resisters of the Heart Mountain Fair Play Committee.

This poster explores the changes that took place in the Japanese American community of Heart Mountain Relocation Center in Wyoming. I will especially investigate the political networks and power relations of the incarceration community. My aim is twofold: on the one hand, to discuss the changes in networks caused by incarceration and dispersal, and on the other, to address some opportunities and challenges presented by the method for the study of history.

Saara Kekki is a Doctoral Candidate at the University of Helsinki. Her dissertation employs network analysis and computer modeling to study the changes in the Japanese American community during and after World War II. She is the author of several articles and co-editor of *North American Studies Crossroads: An Anthology of Finnish Perspectives* (2014) She is the 2011 recipient of the Walter Rundell Award for best dissertation proposal by the Western History Association, being the first non-American to receive the honor. She spent the falls of 2016 and 2017 as a visiting scholar at the University of California, Santa Barbara.

Names as a Part of Game Design

Lasse Hämäläinen
University of Helsinki

Video games often consist of several separate spaces of play. They are called, depending on the speaker and the type of the game, for example *levels*, *maps*, *tracks* or *worlds*. In this paper, the term *level* is used. As there are usually many levels in a game, they need some kind of identifying elements. In some games, levels only have ordinal numbers (*Level 1, Level 2* etc.), but in the other, they also have names.

Names are an important part of game design, at least for three reasons. Firstly, giving names to places makes the imaginary world feel richer and deeper (Schell 2014: 351), improving the gameplay experience. Secondly, name gives the players first impression of the level (Rogers 2014: 220), helping them to perceive the level's structure. And thirdly, level names are needed for discussing the levels. Members of a gaming community often want to share their experiences and emotions of the gameplay. When doing so, it is important to contextualize the events: *in which level did X happen?*

Even though some game design scholars seem to recognize the importance of names, there are very few studies of them. This paper is aimed to fill this blank. I have analyzed level names in Playforia Minigolf, an online game designed in Finland in 2002. The data include names of all the 2,072 levels in the game. The presentation gives an overview of this name system, examining the language, structure and semantics of level names.

References

Playforia (2002). *Minigolf*. Finland: Apaja Creative Solutions Oy.

Rogers, Scott (2014). *Level Up! The Guide to Great Video Game Design*. Chichester: Wiley.

Schell, Jesse (2014). *The Art of Game Design: A Book of Lenses*. CRC Press.

Keywords: onomastics, game studies, game design

Dr. Raluca Fratiloiu	Based on VR content developed by
Department of Communications	UBC's Emerging Media Lab
Okanagan College	and Metanaut VR
Kelowna, BC, Canada	Vancouver, BC, Canada

The Stanley Rhetoric: A Procedural Analysis of VR Interactions in 3D Spatial Environments of Stanley Park, BC

In this paper we propose a closer examination of the key reasons a VR experiential fieldtrip of Stanley Park, British Columbia developed by UBC's Emerging Media Lab in partnership with Metanaut VR is a rhetorically effective discourse. From a rhetorical standpoint, there are wider implications if students and larger audiences are enabled to both create and experience VR content.

Procedural rhetorical analysis in videogames has become a core methodological approach. Procedurality according to Bogost (2007) affects three areas: politics, advertising and learning. Several of these implications have already been investigated. Also, particular attention has been paid to how new media open new possibilities through play and how in turn this creates a renewed interest in digital rhetoric (Daniel-Wariya, 2016). At the same time, procedural rhetoric has been also investigated, at length, in connection to learning through games (Gee, 2007). Learning also has been central in a few studies on VR in education (Dalgarno, 2010). However, there are no specific assessments of procedural rhetoric outcomes of particular VR educational projects.

First, we will outline some theoretical connections that are needed for a rhetorical analysis of virtual reality experiences. Next, we will focus on a rhetorical analysis of preliminary project documents, combined with an interactions analysis in this VR site. Finally, we will propose a critical reflection tool for further consideration once the project will be fully integrated in the classroom.

Starting points: Rhetoric & Procedurality

In order to arrive at how rhetorical analysis is useful in analysing VR experiences, we need to investigate how rhetoric is different in new media relative to traditional texts. A significant amount of work has been done in digital rhetoric and online persuasion (Warnick, 2007). Such

work was partly prompted by a seminal text on the language of new media, in which Manovitch (2002) argued:

> "Traditionally, texts encoded human knowledge and memory, instructed, inspired, convinced, and seduced their readers to adopt new ideas, new ways of interpreting the world, new ideologies. In short, the printed word was linked to the art of rhetoric. While it is probably possible to invent a new rhetoric of hypermedia […] the sheer existence and popularity of hyperlinking exemplifies the continuing decline of the field of rhetoric" (Manovitch, 2002).

Formulating a clear response to such claims is key to subsequently understanding the role rhetoric plays in analysing VR experiences. It's important to note that rhetorical studies outline how any text has potential to be rhetorical. Booth (2004) would probably disagree that new media trigger a decline of rhetoric. A definition of rhetoric by Booth (2004) encompasses the broad terrain rhetoric covers as rhetoric is:

> "… the entire range of resources that human beings share for producing effects on one another: effects [being] ethical (including everything about character), practical (including political), emotional (including aesthetic), and intellectual (including every academic field). It is the entire range of our use of "signs" for communicating effectively or sloppily, ethically or immorally. At its worst, it is our most harmful miseducator – except for violence" (Booth, 2004, p. xi).

For Bitzer (1968) an act of rhetoric is called forth by a "rhetorical situation" that has an effect on audiences. In identifying the exigencies of rhetorical situations, he argues that its most important element is the audience. "Situations are not always accompanied by discourse" (p. 2). Particularly, "it is the situation which calls the discourse into existence" (p. 2). Bitzer (1968) outlines a "theory of situation" as rhetoric is situational, not necessarily with regard to the context of meaning of the situation, which is a "general condition of human communication", or its persuasive character (p. 3). Rhetoric is situational particularly because it can change the audience "in belief or action" (p. 3). In order to clarify "rhetoric-as-essentially-related-to-situation", Bitzer (1968) argues:

> "… A work of rhetoric is pragmatic; it comes into existence for the sake of something beyond itself, it functions ultimately to produce action or change in the world; it performs some task. In short, rhetoric is a mode of altering reality, not by the direct application of energy to objects, but by the creation of discourse which changes reality through the mediation of thought and action" (pp. 3-4).

The definition of a rhetorical situation follows:

> "Rhetorical situation may be defined as a complex of persons, events, objects, and relations presenting an actual or potential exigency which can be completely or partially removed if discourse, introduced into the situation, can so constrain human decision or action as to bring about the significant modification of the exigence" (p. 6).

How does this apply to educational VR experiences? Rhetorical exigencies abound in virtual reality with an educational purpose, as this discourse is usually produced in order to create change, alter belief or reality and move to action.

The use of rhetorical analysis in video games studies is well-established and provides leverage for understanding rhetoric in VR. While all the above-mentioned resources that Booth (2004) mentions are used in video game analysis, rhetorical discourse includes one more important resource - rhetorical procedures. Bogost (2007) introduced the concept of "procedural rhetoric" in *Persuasive Games* as a way of showing how video games' unique modes of persuasion. He argued:

> "…procedural rhetoric is the practice of using processes persuasively, just as verbal rhetoric is the practice of using oratory persuasively and visual rhetoric is the practice of using images persuasively. . . procedural rhetoric is a subdomain of procedural authorship; its arguments are made not through the construction of words or images, but through the authorship of rules of behaviour, the construction of dynamic models" (pp. 28–29).

Thus, in the same way that visual rhetoricians would argue visual rhetoric as the field that adequately deals with the persuasive powers of images, computer games rhetoricians need a field that can explain the type of interactions made possible by this new medium. "Verbal, written, and visual rhetorics inadequately account for the unique properties of procedural expressions. Embodied action is key in assessing these procedural expressions that communicate and represent beyond words and images. Several authors (Gee, 2007; Konzack, 2007; Bates, 2008; Voorhees, 2009) conducted critical rhetorical procedural analyses in commercial or serious games. After connecting several studies on the topic, Paul (2010) concludes the following:

> "Rhetorical analysis offers virtual worlds a perspective for analysis of discourse, especially the procedural, paratextual, and textual discourse that typify virtual worlds.

> The tools of rhetoric help analyze how things work, what they do, and how these kinds of texts interact with each other to shape the context of virtual worlds" (p. 13).

How can this method be adapted to include educational VR experiences? As our case analysis is a virtual field trip, it is important to mention that several studies look into geographical representations of virtual reality (Unwin & Fisher, 2003). Rhetorical procedural analyses of virtual reality environments are however not yet developed.

About the project & case analysis

The goal of this project funded through an Open Education Research grant and developed in Unity3d by University of British Columbia's Emerging Media Lab in partnership with a VR industry company (Metanaut VR) was to develop a proof of concept of an experiential virtual reality field trip that is both immersive and educational. According to the lead authors (Dr. Lock Brown, UBC, Dr. Arthur Gill Green, Okanagan College, Dr. Derek Turner, UBC and Saeed Dyanatkar, Executive Producer at UBC Studios) the project involved a group of at least 15 undergraduate UBC students who produced 70-80% of the work (Green in Emerging Media Lab, 2017). In this process, students had the assistance and support of UBC Studios, UBC Geography, MetanautVR, and the UBC undergraduate society AGDA.

For the purpose of this case analysis, we looked at stated goals outlined by authors in proposals and interim reports. The focus was to understand the motivations of the project and particularly what constitutes the context of this rhetorical situation. We also looked at logs of possible user interactions with the environment, a video walkthrough/tutorial and a wiki developed by the students who contributed to the project. In addition, we experienced the virtual field trip itself.

Let's first investigate the context of the experiences this group of undergraduate students had and how this context speaks about rhetorical exigencies. Students drove the project as they produced content and wrote code for a site that re-created a 3D spatial environment of Stanley Park located in Vancouver, British Columbia. They had no prior experience with either. They had help from professors and industry experts but essentially had an opportunity to dive into

this project and develop a framework from scratch. This was done within an open pedagogy framework. One of the main project goals was "to build the capacity for future projects in the future" (Dyanatkar in Emerging Media Lab, 2017). Assets and workflows are to be released as educational resources and in addition, white papers and a best educational practices document will be published shortly. A report and research article by Dr. Arthur Gill Green will also be available. Current information about the project, now in phase 2 is published on the Geography VR Project Wiki (2017) and the Stanley Park Geography VR Field Trip (2017) website.

The next question is why Stanley Park in particular chosen for this student-led project? Stanley Park, one of the most iconic Canadian destinations is presented in the VR site, as an experiential field trip, featuring educational content as well as 3D spatial environment models of Prospect Point, Beaver Lake, Lumberman's Arch, and the Hollow Tree (BCCampus, 2017). Green (2016) discusses the reasons for choosing this location, one of which was to "lower barriers accessing field locations" (para. 2). In this sense, he argues, VR has an unexplored potential for creating interactive content around landmark sites that would otherwise not be available to many as field trip experiences. Also, "there has been very little work on best practices for linking pedagogically founded learning goals to VR resources" which led to the impetus for an "experimental educational project" (para 2). So, how does this stated goal speak to a change in discourse in the way field trips can be conducted? We argue it does that in at least three ways: (1) VR field trips may help remove financial and logistic barriers for numerous students, (2) VR field trips prompt audiences to experience and interact with educational content that would otherwise be accessible only in more traditional formats, and (3) VR field trips can encourage free exploration of and revisits to a designed spatial environment whereas field trips often entail experts leading students through one visit to a location.

Given the limited abilities of educational institutions to lead and develop field trips, alternative approaches are needed. According to the authors, this project may fill in an important gap, as new technologies such as VR and augmented reality (AR) have the potential to provide more opportunities for experiential learning.

In addition to these core exigencies, this particular case also reveals how VR and AR experiences enable one to critically assess various histories of a place and develop a layered understanding of the implications of travelling to a location. When going in the site to Prospect Point for example we are welcome by the following audio content:

> "Welcome to Prospect Point. This famous viewpoint is one of the most popular tourist attractions in Stanley Park. […]. This forest surrounding you is an ancient wilderness… or is it? The trees you see around you certainly mirror the appearance of forests in pre-industrial times" (Emerging Media Lab, 2016a).

The audio content further invites the traveller to "look closer", contemplate how this natural environment is constantly changing and think about the role deliberate planning and policies had in shaping the way the park is today.

In the VR experience of Stanley Park, one can experience the landscape but also begin to engage with a more nuanced history as visitors/students/users have opportunities to explore the complex history of this impressive location that was once home to Burrard, Musqueam and Squamish First Nations people (City of Vancouver, 2017). When reaching Lumberman's Arch travellers/students/visitors are asked to consider "another history… one that is not easy to look at" (Emerging Media Lab, 2016b). The accompanying audio says:

> "The history of Stanley Park is inseparable from Canada's history of the marginalization of its indigenous people. When Lord Stanley declared this part to be "for people of all creeds and customs", a few restrictions applied. He neglected to include the local first nations communities who already lived here. So today - - it's hard to spot the evidence that this was once a thriving village. Yet another example of how indigenous history from Stanley Park was erased" ("Lumberman's Arch" UBC student content, 2016).

The audio recording goes further into issues with these two parallel histories of colonialism, on one hand and indigenous heritage, on the other and explores current problems with consumerist tourism practices that create a "semblance of First Nations in the park, despite the fact that authentic culture has been removed" (Emerging Media Lab, 2016b).

The above are just some examples speaking to the ethos conveyed via audio content about the park itself. A longer and more detailed analysis should include other examples that lend themselves on a rhetorical and textual analysis of the audio content accompanying the history

of Lion's Gate Bridge or Hollow Tree. For now, let's shift the focus on how rhetorical procedures add to the experience of visiting these VR locations.

From a procedural standpoint several types of interactions are possible in this VR experience of Stanley Park. We have already referred to examples audio content rendered via "cassette players" which can be accessed and stopped anytime. Other interactions in the environment are: walk/teleport, hand interaction with the environment, viewing a map, watching 360 videos, info panels and photospheres of other locations. Reading through the project logs of potential user experiences, we can identify some of the goals attached to these procedures. Walking and teleporting allow users to explore the environment at their own pace but also to jump from one area to another. Interactions with the environment allow users to experience the park like they are physically there. One can interact with the geology or plant models and learn about the place in ways that textbooks do not allow one to learn. Using one's hands allows users to interact and play with the environment. Maps allow users to understand the location and destination and get a bigger picture of the destination. 360 videos, cassette players, info panels and photospheres allow users to learn more about the place and make sure they learn about the key features and do not miss important points about the experience. They also serve to explain the material in more depth in an interactive fashion.
The combination of audio content, text featured via info panels and interactions with objects in the environment makes this VR experiential field trip rhetorically effective from a procedural standpoint. Through the VR experience one is able to more tangibly interact with objects that would otherwise be abstract or one would read about. Also, the range of interactions renders the experience "realistic" which is in line with the stated goal for the project to create a "realistic field-trip environment" where one is able to interact with different parts of the environment and learn about the site, not just walk through and admire the scenery (Peter Kao in Emerging Media Lab, 2017).

In addition, the VR field trip prompts one to consider the complicated history of this location and analyse rhetorical implications that are not immediately obvious. Stanley Park was "created with the intention of showcasing the wild coastal forests" but it is actually "far from a wild place" (Stanley Park History, UBC student content, 2016). As everything in the park has been

managed students/users/visitors are asked to consider how “our concept of nature changes” and how we interact with it (Emerging Media Lab, 2016c).

Several tools are available for analysing video games as media texts. The most cited/used one is probably the one developed by Consalvo and Dutton (2006). Several other authors argue for the need to develop more connections between critical communication, rhetorical theory and game studies (Voorhes, 201). Some authors argue for game-based pedagogies in writing classrooms (Shultz Colby, 2017). Here we begin to contribute to a pedagogy of teaching writing about VR experiences via a combination of textual and rhetorical methods.

Based on our analysis of this particular educational VR experience, we propose a teaching tool that can used in the classroom. There is a fair bit of literature that discusses how games offer multiple opportunities for learning. One such opportunity is writing about games. Developing writing reflective assignments helps in solidifying the knowledge and reflecting on the experience. Combining elements of Consalvo and Dutton’s (2006) list of elements that lead to a comprehensive textual analysis of video games with aspects of the rhetorical situation and procedural rhetoric can lead to a set of open-ended questions that students could think through when interacting with a VR environment:

1. What do you think is the goal of this experience?
2. How does it feel to be in this virtual place?
3. How does the tutorial work for you? Rate its effectiveness:
   1 (low)_ _ _ _5 (high)
4. Can you interact with the objects described in the tutorial? Rate your interactions:
   1 (low)_ _ _ _5 (high)
5. What is your favourite action/interaction and why? What does each do? (*Specific examples like orbs, walking, teleporting can be added.)*
6. What are other complementary rhetorical devices (music, text, images, etc). Do they enhance or deter from the virtual experience? Explain why.
7. Do you need paratexts (lectures, online materials, etc) to make sense of the experience or does the VR experience make sense on its own? Explain what materials are useful/not useful.
8. What have you learnt from exploring this site?
9. What do you wish this experience did in addition to what it currently does?
10. Would you come back? Why or why not? Would you advise someone else to experience this fieldtrip?

Conclusion and relevance

Beyond this specific VR case, we need to examine rhetoric in VR with an educational goal in order to assess pedagogical effectiveness and success with audiences. Considering the rhetorical impact of VR's affordances may allow developers to enhance the potential of meaningful interactions with students and other users. This case analysis reveals that meaningful interactions are possible in VR designed with an educational purpose. Also, it is a case that demonstrates the potential to develop VR projects in imaginative ways with less resources than one expects whilst making them widely available. In addition, a VR educational experience gives a sense of a place but also enhances it through added functionality. Supplemental materials and paratetxts can be very useful. VR experiences allow for a range of objects to be already embedded in the experience itself and contribute to learning. In addition, the project reveals the potential to lower cost barriers to field trips but also to lower cost barriers to creating VR educational content. Enlisting student work with the goal of creating such educational materials provides ample opportunities for applied learning across disciplines. Also, the creation of open education VR resources creates possibilities for wider audiences to engage with this new technology. Consequently, this case analysis may open up new possibilities for investigating how students/users derive meaning from interacting in these environments and continue a dialogue between several connected areas of education and VR, games and pedagogy, games and procedural rhetoric.

References

BCCampus. (2017, May 10). *Virtual reality and augmented reality field trips funded by OER grants.* Retrieved from BCCampus: https://bccampus.ca/2017/05/10/virtual-reality-and-augmented-reality-field-trips-funded-by-oer-grants/.

Bitzer, L. (1968). The rhetorical situation. *Philosophy and Rhetoric*, 1, pp. 1-14.

Bogost, I. (2007). *Persuasive Games: The Expressive Power of Videogames.* Cambridge: MA: MIT Press:.

Booth, W. C. (2004). *The Rhetoric of Rhetoric: The Quest for Effective Communication.* Malden:MA.: Blackwell Publishing.

City of Vancouver. (2017). *The History of Stanley Park.* Retrieved from City of Vancouver: http://vancouver.ca/parks-recreation-culture/stanley-park-history.aspx.

Consalvo, M. and Nathan Dutton (2006). Game analysis: Developing a methodological toolkit for the qualitative study of games. *Game Studies: The international journal of computer game research,* Vol 6, Issue1, December 2006. Retrieved from: http://gamestudies.org/0601/articles/consalvo_dutton.

Dalgarno, L. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology*, Vol 41, No 1, pp. 10-32.

Daniel-Wariya, J. (2016). A Language of Play: New Media's Possibility Spaces. *Computers and Composition*, 40, pp. 32-47.

Emerging Media Lab (2016a). Welcome to Prospect Point. Vancouver, BC, Canada.

Emerging Media Lab (2016b). Lumberman's Arch. Vancouver, BC, Canada.

Emerging Media Lab (2016c). Stanley Park History. Vancouver, BC, Canada.

Emerging Media Lab (2017). *Stanley Park Geography VR Field Trip.* Retrieved from The University of British Columbia, Vancouver Campus: http://eml.ubc.ca/projects/geography-vr/

Emerging Media Lab (Director). (2017). *Geography Virtual Reality Fieldtrip* [Video]. Retrieved from: http://eml.ubc.ca/projects/geography-vr/

Gee, J. P. (2003). *What video games have to teach us about language and literacy.* New York: Palgrave Macmillan.

Gee, J. P. (2007). *What Video Games Have to Teach Us About Learning and Literacy. Second Edition: Revised and Updated Edition.* New York: St. Martin's Griffin.

Green, A. (2016, July 9). *Virtual Reality, 3D Spatial Environments, and Environmental Education.* Retrieved from Greengeographer.com: http://greengeographer.com/photogrammetry-and-oer-field-trips/.

Manovitch, L. (2002). *The language of new media.* Cambridge, MA: MIT Press.

Paul, C. (2010). Process, Paratexts, and Texts: Rhetorical Analysis and Virtual Worlds. *Journal of Virtual Worlds Research*, Volume 3, Number 1, pp. 4-17.

Shultz Colby, R. (2017). Game-based Pedagogy in the Writing Classroom. *Computers and Composition*, 43, pp. 55-72.

UBC Geography students. (2017, December). *Geography VR Project Wiki*. Retrieved from https://sites.google.com/view/ubcgeovr/home.

Unwin, D., & Fisher, P. (2003). *Virtual Reality in Geography.* eBook: Taylor & Francis.

Voorhes, G. (2012). Discursive Games and Gamic Discourses. *Futures of Communication*, August, Vol 1, article 3, pp. 1-21.

Warnick, B. (2007). *Rhetoric Online: Persuasion and Politics on the World Wide Web.* New York: Peter Lang.

Cultural heritage collections as research data

Cultural heritage materials held in institutional collections are crucial sources of evidence for many disciplines, ranging from history and literature to anthropology and art. They are also the subjects of research in their own right – encompassing their form, their history, and their content, as well as their places in broader assemblages like collections and ownership networks. They can be studied for their unique and individual qualities, as Neil McGregor demonstrated in his *History of the World in 100 Objects,* but also as components within a much larger quantitative framework.

Large-scale research into the history and characteristics of cultural heritage materials is heavily dependent on the availability of collections data in appropriate formats and sufficient quantities. Unfortunately, this kind of research has been seriously limited, for the most part, by lack of access to suitable curatorial data. In some instances this is simply because collection databases have not been made fully available on the Web. This is particularly the case with art galleries and some museums. Even where databases are available, however, they often cannot be downloaded in their entirety or through bulk selections of relevant content. Data downloads are frequently limited to small selections of specific records.

Collections data are often available only in formats which are difficult to re-use for research purposes. In the case of libraries, the only export formats tend to be proprietary bibliographic schemas such as EndNote or RefCite. Even where APIs are made available, they may be difficult to use or limited in their functionality. CSV or XML downloads are relatively rare. Data licensing regimes may also discourage re-use, either by explicit limitations or by lack of clarity about terms and conditions.

Even where researchers are able to download usable data, it is very rare for them to be able to feed back any cleaning or enhancing they may have done. The cultural heritage institutions supplying the data may be unable or unwilling to accept corrections or improvements to their records. They may also be suspicious of researchers developing new digital services which appear to compete with the original database.

As a result, there has been a significant disconnect between curatorial databases and researchers, who have struggled to make effective use of what is potentially a very rich source of computationally usable evidence. One important consequence is that re-use of curatorial data by researchers often focuses on the data which are the easiest to obtain. The results are neither particularly representative nor exhaustive, and may weaken the validity of the conclusions drawn from the research.

Some recent “collections as data” initiatives (such as collectionsasdata.github.io) have started to explore approaches to best practice for “computationally amenable

collections", with the aim of "encouraging cultural heritage organizations to develop collections and systems that are more amenable to emerging computational methods and tools". Under the auspices of the Library of Congress and the Institute of Museum and Library Services, the Collections as Data programme "aims to foster a strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research" (Always Already Computational 2017). One of the drivers for this initiative is the perception that, as Miriam Posner argues, "Libraries and archives [and museums] are increasingly making their materials available online, but, as a general rule, these materials aren't of much use for computational purposes" (Posner 2017).

This paper focuses on three case studies of projects which are addressing these issues. The first project is "Collecting the West", in which Western Australian researchers are working with the British Museum to deploy and evaluate the ResearchSpace software, which is designed to integrate heterogeneous collection data into a cultural heritage knowledge graph in a Linked Data environment. The second project is HuNI – the Humanities Networked Infrastructure – which has been building a "virtual laboratory" for the humanities by reshaping collections data into semantic network graphs. The third project – "Reconstructing the Phillipps Collection", funded by the European Union under its Marie Curie Fellowships scheme – involved combining collections data from a range of digital and physical sources to reconstruct the histories of manuscripts in the largest private collection ever assembled.

To make services like these possible, collections data need to be made available in certain ways and under certain conditions. Recommendations for best practice, at the moment, tend to be focused mostly on processes and procedures, encompassing download formats, licensing, and availability in particular (Fitzpatrick 2017). These are undoubtedly important; having collections data easily accessible in bulk on the Web, under a Creative Commons licence which permits free reuse, is essential. Download formats are more debatable: APIs are not necessarily the best approach, given that their use is likely to require a significant level of technical expertise (Tauberer 2014). XML dumps and CSV files are easier to use, but may not contain all the elements in the source database.

As the interest of researchers in reusing collections data continues to grow, however, cultural heritage institutions increasingly need to start looking beyond simply making their data available for bulk downloading or via an API. One of the major use cases is to link together data from different institutions, without diminishing the semantic richness, in order to ask questions on a larger scale. At the moment, researchers are having to do much of this work themselves. This raises two important questions: should institutions help this process, and what kind of infrastructure might be built as a result?

The prominence of Linked Data in the solutions being adopted by researchers strongly suggests that institutions should make their data available in formats suitable for incorporation into Linked Data environments. While many institutions

might not yet see a 'business case' for this approach, others like the British Library and the British Museum have already followed this route. Making available an RDF version of a relational database would be a significant contribution. But even embedding into that database identifiers which point to widely-used Linked Data ontologies and vocabularies like VIAF, GeoNames and Wikidata would be valuable. So too would taking a critical look at ways of improving the computational value of ownership and provenance data in these records. Enabling researchers and curators to annotate and add to the data is also emerging as an important requirement.

Beyond this, though, lies the wider landscape of digital infrastructure. The *Santa Barbara Statement on Collections as Data* (2017) observes that "Working toward interoperability entails alignment with emerging and/or established community standards and infrastructure." At present, the Linked Data landscape is largely being built by research groups rather than cultural institutions, which still tend to focus on their own collections. In this context, an initiative like "Linked Pasts", which has emerged from the Pelagios Commons, is an important development, offering a vision of joining up disparate Linked Data projects in the humanities to create a "wider ecosystem" (Grossner and Hill 2017).

As long as these kinds of initiatives remain tied to research projects, their future sustainability will be reliant on the uncertainty of grant funding. Collecting institutions should look closely at them as outcomes of the reuse of collections data, and consider seriously the value of partnerships with the researchers involved. They should recognize that there is a growing group of researchers who do not simply want to search or browse a collections database. There is an increasing demand for access to collections data for downloading and re-use, in suitable formats and on non-restrictive licensing terms. In return, researchers will be able to offer enhanced and improved ways of analyzing and visualizing data, as well as correcting and amplifying collection database records on the basis of research results. There are significant potential benefits for both sides of this partnership.

Bibliography

Always Already Computational: Library Collections as Data. 2017. https://collectionsasdata.github.io/

Burrows, Toby. 2017. "The History and Provenance of Manuscripts in the Collection of Sir Thomas Phillipps: New Approaches to Digital Representation," *Speculum* 92 S1:S39-S64

Burrows, Toby, and Deb Verhoeven. 2015. "Aggregating Cultural Heritage Data for Research Use: The Humanities Networked Infrastructure (HuNI)." In *Metadata and Semantics Research, 9th Research Conference, MTSR 2015, Manchester, UK, September 9–11, 2015: Proceedings*, ed. Emmanouel Garoufallou, Richard J. Hartley, Panorea Gaitanou (Communications in Computer and Information Science, 544), 417-423. Cham: Springer.

Fitzpatrick, L. Kelly. 2017. “Shared Practices in Museum Open Collections Data,” *Medium*, February 22. https://medium.com/berkman-klein-center/shared-practices-in-museum-open-collections-data-72e924c4849a

Flanders, Julia. 2014. “Rethinking Collections.” In *Advancing the Digital Humanities*, edited by Paul Longley Arthur and Katherine Bode, 163-174. London: Palgrave MacMillan.

Grossner, Karl, and Timothy Hill. 2017. “From Linking Places to a Linked Pasts Network.” http://kgeographer.com/pubs/LinkedPastsNetwork_7Dec.pdf

Hyvönen, Eero. 2012. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. San Rafael, CA: Morgan & Claypool.

MacGregor, Neil. 2010. *A History of the World in 100 Objects*. London: Penguin.

Posner, Miriam. 2017. “Actually Useful Collection Data: Some Infrastructure Suggestions.” In *Always Already Computational: Library Collections as Data: National Forum Position Statements.*
https://github.com/collectionsasdata/collectionsasdata.github.io/raw/master/aac_positionstatements.pdf

Santa Barbara Statement on Collections as Data. 2017.
https://collectionsasdata.github.io/statement/

Tauberer, Joshua. 2014. “Bulk data or an API?” https://opengovdata.io/2014/bulk-data-an-api/

Digital Humanities in the Nordic Countries 2018

Big Data and the Afterlives of Medieval and Renaissance Manuscripts: the "Mapping Manuscript Migrations" Project

Abstract

Tens of thousands of European medieval and Renaissance manuscripts have survived until the present day. As the result of changes of ownership over the centuries, they are now spread all over the world, in collections across Europe, North America, Asia and Australasia. They often feature among the treasures of libraries, museums, galleries, and archives, and they are frequently the focus of exhibitions and events in these institutions. They provide crucial evidence for research in many disciplines, including textual and literary studies, history, cultural heritage, and the fine arts. They are also objects of research in their own right, with disciplines such as paleography and codicology examining the production, distribution, and history of manuscripts, together with the people and institutions who created, used, owned, and collected them.

Over the last twenty years there has been a proliferation of digital data relating to these manuscripts, not just in the form of catalogues, databases, and vocabularies, but also in digital editions and transcriptions and – especially – in digital images of manuscripts. Overall, however, there is a lack of coherent, interoperable infrastructure for the digital data relating to these manuscripts, and the evidence base remains fragmented and scattered across hundreds, if not thousands, of data sources.

The complexity of navigating multiple printed sources to carry out manuscript research has, if anything, been increased by this proliferation of digital sources of data. Large-scale analysis, for both quantitative and qualitative research questions, still requires very time-consuming exploration of numerous disparate sources and resources, including manuscript catalogues and databases of digitized manuscripts, as well as many forms of secondary literature. As a result, most large-scale research questions about medieval and Renaissance manuscripts remain very difficult, if not impossible, to answer.

The "Mapping Manuscript Migrations" project, funded by the Trans-Atlantic Platform under its Digging into Data Challenge for 2017-2019, aims to address these needs. It is led by the University of Oxford, in partnership with the University of Pennnsylvania, Aalto University in Helsinki, and the Institut de recherche et d'histoire des textes in Paris. The project is building a coherent framework to link manuscript data from various disparate sources, with the aim of enabling searchable and browsable semantic access to aggregated evidence about the history of medieval and Renaissance manuscripts.

This framework is being used as the basis for a large-scale analysis of the history and movement of these manuscripts over the centuries. The broad research questions being addressed include: how many manuscripts have survived;

where they are now; and which people and institutions have been involved in their history. More specific research focuses on particular collectors and countries.

The paper will report on the first six months of this project. The topics covered will include the sources of data which are being combined, and the data modeling being carried out to link disparate data sources within a Linked Data environment. Data from four different sources are being combined initially, including three relational databases – the Schoenberg Database of Manuscripts, Bibale, and Medium – and a catalogue built from TEI-encoded documents (Medieval Manuscripts in Oxford Libraries). Transforming these into RDF, mapping them to an ontology derived from CIDOC-CRM and FRBRoo, and linking them to external identifiers from VIAF, Wikidata, and GeoNames, raise various issues and challenges.

The paper will also report on the new digital platform being developed, and how it is being informed by specific research questions derived from an analysis of the literature and from discussions with a focus group of manuscript researchers. Requirements for visualizing and navigating the data will be discussed, drawing on work previously done by a project to analyse the history of the manuscript collection of Sir Thomas Phillipps.

Bibliography

Burrows, Toby (2017) "The History and Provenance of Manuscripts in the Collection of Sir Thomas Phillipps: New Approaches to Digital Representation," *Speculum* 92 S1, S39-S64

Da Rold, Orietta and Marilena Maniaci (2015) "Medieval Manuscript Studies: a European Perspective", in: *Writing Europe, 500-1450: Texts and Contexts* (Essays and Studies, 68), ed. Aidan Conti, Orietta Da Rold & Philip Shaw (Cambridge: D. S. Brewer, 2015), pp. 1-24

Hyvönen, Eero, Jouni Tuominen, Miika Alonen and Eetu Mäkelä (2014) "Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets", in: *The Semantic Web: ESWC 2014 Satellite Events* (Lecture Notes in Computer Science, 8798) (Berlin: Springer-Verlag, 2014), pp. 226-230

Scase, Wendy (2009) *Applying Semantic Web Technologies to Medieval Manuscript Research: European Science Foundation Exploratory Workshop Report*. Strasbourg: European Science Foundation, 2009.

The New Face of Ethnography: Utilizing Facebook as an Alternative Study Site

The transfer of children has been a traditional means of survival for both families and children in cultures throughout history as an informal, non-permanent arrangement. Plenary adoption, as practiced in the United States since the middle 20th century is very different in that it requires the permanent, legal severing and reassigning of kinship ties. This is not natural but the practices surrounding it as well as the stereotypes, tropes and stories weave themselves into society and make it ordinary and normalize the practice.

Adoption practices have changed over time and the understanding of adoption both publicly and privately have changed as well. Popular and widespread tropes surrounding adoption do not adequately cover the emotional and lived experience of those involved in the process but they do create a rubric to explain how one should feel about adoption and how one should embody the experience. The change from closed to open adoption and the introduction of international and interracial adoption prompted a change in the American definition of "family" and the way adoption was conducted and understood as a social institution. The narratives of those intimately involved in the adoption process begin to reflect those changes and incorporate them into the adoption story.

Online ethnography presents both benefits and complications. On the plus side, online communities break down the walls of networks, creating digitally mediated social spaces. The Internet provides a platform for social interactions where real and virtual worlds shift and conflate. Social interactions in virtual environments present another option for social researchers and offer significant advantages for data collection, collaboration, and maintenance of research relationships. For some research subjects, such as members of the adoption triad, locating target groups presents challenges for domestic adoption researchers. Online groups such as Facebook pages dedicated to specific members of the adoption triad offer a resolution to this challenge, acting as self-sorted focus groups with participants eager to provide their narratives and experiences.

Ethnography involves understanding how people experience their lives through observation and non-directed interaction, with a goal of observing participants' behavior and reactions on their own terms; this can be achieved through the presumed anonymity of online interaction. Electronic ethnography provides valuable insights and data; however, on the negative side, the danger of groupthink in particular Facebook communities can both attract and generate homogeneous experiences regarding adoption issues.

Adoption narratives are not simply a matter of constructing personal stories or chronicling an event, they are a reflection of social understanding and belief systems. This project addresses the silences, erasures, and peripheralization that exist in the dominant national adoption

narrative. Normative narratives of adoption contextualize it as providing a home for homeless children and children for childless families. This frame prevails in public discussions, but elides less positive and more complex aspects of adoption: the emotional situation of birth mothers, the family secrets kept from adoptees, and fear of adoptive parents, etc.

I bring the voices of the adoptive parents, birth parents and adoptees to the forefront of the discussion to contest and destabilize normative adoption narratives and create an alternative adoption narrative that acknowledges the ruptures that must take place before adoption is possible. I provide an analysis of personal narratives from birth/first mothers, adoptive parents and adoptees to expose how adoption has defined members of the primary groups within the constellation of adoption and how their performance of identity both conforms and pushes back against popular tropes and "common knowledge" about adoption.

My project works toward a (re)formation of adoption narratives that capture experiences not recognized in how adoption is presented socially and emotionally. I (re)create an adoption narrative that amplifies the voices that have historically been pushed to the background and dismissed. This alternative adoption narrative will problematize the win-win solution and offer a different perspective on the mass transfer of children in the United States. To explore these issues I rely on more than 200 narratives submitted by members of the adoption triad through a completely anonymous online survey.

My anthropological research, as it relates to the constructivism of narratives that inform identity, is novel and methodologically innovative. My analysis of the unique and contingent nature of the adoptee-adopter-biological parent connection turns the performative "structures of sentiment" that have informed the modern adoptive process on their heads. My methodology reflects the turn toward social media communities as a source of data.

The paradigm shift my work offers is necessary to understand the lasting impact that adoption has on members of the adoptive triad beyond childhood. Adoption is not simply a single event, but instead follows the participants through life and gathers new meaning as greater understanding of the self and relationship to others is exposed. My work, specifically the anthropological inquiry of familial roles, performativity and narrative construction as a proxy for self-identity, is part of a new theorization of the ways context and identity matters within the institution of adoption.

Keywords: Adoption, Ethnography, Social Media, Narrative

Research in Nordic literary collections: What is possible and what is relevant?

Mads Rosendahl Thomsen, Mats Malm and Kristoffer Laigaard Nielbo
Aarhus University, Gothenburg University, University of Southern Denmark

There are a growing number of digital literary collections in the Nordic countries that make the literary heritage accessible and have great potential for research that takes advantage of machine readable texts. These collections range from very large collections such as the Norwegian Bokhylla, medium-sized collections such as the Swedish Litteraturbanken and the Danish Arkiv for Dansk Litteratur, to one-author collections, e.g. the collected works of N.F.S. Grundtvig. In this presentation we will discuss some of the obstacles for a more widespread use of these collections by literary scholars and present outcomes of a series of seminars – UCLA 2015, Aarhus 2016, UCLA 2017 – sponsored by the Fondation Maison des sciences de l'homme courtesy of a grant from the Andrew Carnegie Mellon Foundation.

We find that there are two important thresholds in the use of collections:
1) The technical obstacles for collecting the right corpora and applying the appropriate tools for analysis are too high for the majority of researchers working in literary studies. While much has been done to advance the access to works, differences in formats and metadata make it difficult to work across the collections. Our project has addressed this issue by creating a Nordic github repository for literary texts, CLEAR, which provides cleaned versions of Nordic literary works, as well as a suite of tools in Python.

2) The capacity to combine traditional hermeneutical approaches to literary studies with computational approaches is still in its infancy despite numerous good studies from the past years, e.g. by Stanford Literary Lab, Leonard and Tangherlini and Ted Underwood. We have worked to bring together scholars with great technical prowess and more traditionally trained literary scholars in a series of seminars, to generate projects that are technically feasible and scholarly relevant. The process of expanding the methodological vocabulary of literary studies is complicated and requires significant domain expertise to verify the outcome of computational analyses, and conversely, openness to work with results that cannot be verified by close readings.

In this paper, we will present a Nordic repository for literature and discuss the challenges and choices involved with this. We shall also present a test-case on how thematic variation and readability can provide new perspectives on Swedish and Danish literature, and how prototypes of a dashboard for accessing the repository and data on textual coherence can improve scholarship on literature.

Literature

Algree-Hewitt, Mark et al. 2016. ”Canon/Archive. Large-scale Dynamics in the Literary Field.” *Stanford Literary Lab* Pamphlet 11.

Heise, Ursula. 2017. “Comparative literature and computational criticism: A conversation with Franco Moretti.” *Futures of Comparative Literature: ACLA State of the Discipline Report*. London: Routledge, 2017.

Leonard, Peter and Timothy R. Tangherlini. 2013. “Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research”. *Poetics* 41(6): 725-749.

Thomsen, Mads Rosendahl et al. 2015. “No Future without Humanities.” *Humanities* 1.

Underwood, Ted. 2013. *Why Literary Period Mattered*. Stanford: Stanford University Press.

Using ArcGIS Online and Story Maps to visualize spatial history: The case of Vyborg

A short paper proposal

Historical GIS (HGIS) or spatially oriented history is a field that uses geoinformatics to look at historical phenomena from a spatial perspective. GIS tools are used to visualize, manage and analyze geographical data. However, the use of GIS tools requires some technical expertise and ready-made historical spatial data is almost non-existent, which significantly reduces the reach of HGIS. New tools such as Esri's ArcGIS Online (AGOL) should make spatially oriented history more accessible.

AGOL allows making internet visualization of maps and map layers created with Esri's more traditional GIS desktop program ArcMap. In addition, Story Maps do not necessarily require GIS expertise to create. While the analyses are possible using desktop GIS software, tools such as AGOL facilitate the creation of more powerful visualizations in an easily digestible narrative format for the public.

I will demonstrate the use of Story Maps to represent spatial change in the case of the city of Vyborg. Vyborg is a suitable case study for Historical GIS research since there are significant amounts of archival data available from early modern period. Oldest maps date back to 1630s, while in later centuries maps for both civilian and military use are widely available. The town has also been extremely important in Finnish context, as it has been a heavily fortified garrison town, a center of culture, as well as a hub of commerce and industry. The population of Vyborg was also unusually diverse in Finnish context, which adds another layer of complexity to the spatial history of the town. However, there has not been any spatially oriented historical research looking into Vyborg. Earlier research on e.g. segregation of different ethnic groups has only examined it from an aspatial perspective, using rough district-based summaries. GIS, on the other hand, allows analyses that are far more detailed. The availability of accurate population registers in the 19th century makes it possible to track demographic composition even at sub-block level, while Story maps make their visualization easier.

The city of Vyborg lies in Russia near the Finnish border. A small town grew near the castle founded by Swedes in 1293. Vyborg was granted town privileges in 1403, and later in the 15th century, it became one of the very few walled towns in Kingdom of Sweden. The town was located on a hilly peninsula near the castle. Until 17th century the town space was 'medieval' i.e. irregular. The town was regulated to conform to a rectangular street layout in 1640s. The first surviving maps of the town are from 1638—1640 showing both the pre- and post-regulation street layout. I show the similarities between old and new town plans by superimposing them on a map. Simple superimpositions help to map the common elements in pre- and post-regulation town layouts.

The Swedish period ended when the Russians conquered Vyborg in 1710. Vyborg became a provincial garrison town and administrative center. Later, when Russia conquered rest of Finland in 1809, the province of Vyborg (aka 'Old Finland') was added to the Autonomous Grand Duchy of Finland, a part of the Russian empire. During the 19th century Vyborg became increasingly

important trade and industrial center, and the population grew rapidly. I map expanding urban areas in the 19th and early 20th century using old town plans and population statistics.

Another perspective to the changing town space is the growth of fortifications around Vyborg. As the range of artillery grew, the fortifications were pushed further and further outside the original town. I use story maps to show the position of fortifications of different eras by placing them in the context of terrain. I also employ viewshed analyses to show how the fortifications dominate the terrain around them. Earlier research has noted the strategically advantageous location of Vyborg and its surroundings, as well as the expansion of the fortifications as the range of guns grew. GIS tools and AGOL, however, make spreading this information much easier.

Keywords: AGOL, Story Maps, Historical GIS, urban history, spatial analysis

Digital Humanities in the Nordic Countries / 7–9 March 2018

Abstract of a presentation
Hanna-Riikka Roine, PhD
Helsinki Collegium for Advanced Studies
hanna.roine@helsinki.fi
February 2, 2018

The Future of Narrative Theory in the Digital Age?

As it has often been noted, digital humanities are to be understood in plural. It seems, however, that quite as often they are understood as the practice of introducing digital methods to humanities, or a way to analyse "the digital" within the humanist framework. This presentation takes a slightly different approach, as its aim is to challenge some of the underlying biases within a humanist field, narrative theory, through the properties of today's computational environment.

My presentation starts from the fact that the ancient technology of storytelling has now become enmeshed in a software-driven environment. This development (and "digital turn", in general) has so far mostly escaped the attention of narratologists, although it has had profound effects on the affordances and environments of storytelling in media. For example, computational media not only has the potential to simulate (or "transmediate") all artistic media, but also differs fundamentally from verbal language in its structure and strategies.

For its part, narrative theory originates from literary criticism and bases its concepts and understanding of narrative in media mostly on printed works. While few trends with a more broadly defined base are emerging (e.g. the project of "transmedial narratology"), the analysis of verbal narrative structures and strategies from the perspective of literary theory remains the primary concern of the field (see Kuhn & Thon 2017). Furthermore, the focus of current research is quite medium-specific, while various phenomena studied by narratology (e.g. narrativity, worldbuilding) are agreed to be medium-independent.

In my presentation, I briefly illustrate the underlying biases of current narrative theory through the properties of computational media. As software-driven, conditional, and process-based, storytelling in computational environments is not so much about disseminating a single story, but rather about multiplication of narrative, centring upon the underlying patterns on which varied instantiations can be based. Furthermore, they challenge the emphasis on fixed media content and author-controlled model of transmission. (See e.g. Murray 1997 and 2011; Bogost 2007, Hayles 2012, Manovich 2013, Jenkins et al. 2013.)

I argue that if the biases are not addressed, narrative theory cannot genuinely deal with the "new norm" represented by computational media. The norm is "new" compared to the prototypical narrative developed in the study of literary fiction. For this reason, narratologist Brian McHale has recently predicted that narrative theory "might become divergent and various, multiple *narratologies* instead of one – a separate narratology for each medium and intermedium" (2016, original emphasis).

In my view, such a future fragmentation of the field would only diminish the potential of narrative theory. Instead, the various theories could converge or hybridize in a similar way that contemporary media has done – especially in the study of today's transmedia which is hybridizing both in the sense of

content being spread across media and in the sense of media being incorporated by computer and thus, acquiring the properties of computational environments.

Narrative theory can, thus, truly contribute to the study of storytelling practices and strategies in contemporary computational media, but various biases underlying its toolkit must be genuinely addressed first. The need for this is urgent not only because "narratives are everywhere", but also because the old traditional online/offline distinction has begun to disappear.

References

Bogost, Ian. 2007. *Persuasive Games: The Expressive Power of Videogames*. Cambridge, Ma: The MIT Press.

Hayles, N. Katherine. 2012. *How We Think: Digital Media and Contemporary Technogenesis*. Chicago: Univ. of Chicago Press.

Jenkins, Henry, Sam Ford, and Joshua Green. 2013. *Spreadable Media: Creating Value and Meaning in a Networked Culture*. New York: New York Univ. Press.

Kuhn, Markus and Jan-Noël Thon. "Guest Editors' Column. Transmedial Narratology: Current Approaches." *NARRATIVE* 25:3 (2017): 253–255.

Manovich, Lev. 2013. *Software Takes Command: Extending the Language of New Media*. New York and London: Bloomsbury.

McHale, Brian. "Afterword: A New Normal?" In *Narrative Theory, Literature, and New Media: Narrative Minds and Virtual Worlds*, edited by Mari Hatavara, Matti Hyvärinen, Maria Mäkelä, and Frans Mäyrä, 295–304. London: Routledge, 2016.

Murray, Janet. 1997. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. New York: The Free Press.

———. 2011. *Inventing the Medium. Principles of Interaction Design as a Cultural Practice*. Cambridge, Ma: The MIT Press.

DH-Nordic conference, Helsinki, March 7-9, 2018

Charting the 'Culture' of Cultural Treaties: Digital Humanities approaches to the history of international ideas

Benjamin G. Martin
Department of History of Science and Ideas, Uppsala University
in collaboration with HUMlab, Umeå University

Cultural treaties are the bi-lateral or multilateral agreements among states that promote and regulate cooperation and exchange in the fields of life we call cultural or intellectual. Pioneered by France just after World War I, this type of treaty represents a distinctive technology of modern international relations, a tool in the toolkit of public diplomacy, a vector of "soft power." A comparative examination of these treaties can identify their role in the history of public diplomacy and in the broader history of culture and power in the international arena. But these treaties can also serve as sources for the study of what the historian David Armitage has called "the intellectual history of the international." In this project, I use digital humanities methods as one part of a multi-method effort to use cultural treaties as a historical source with which to explore the emergence of a hegemonic concept of culture in twentieth century international society. Specifically, the project will investigate the hypothesis that the culture concept, in contrast to earlier ideas of civilization, played a key role in the consolidation of the post-World War II international order.

I approach the topic by charting how concepts of culture were given form in the system of international treaties between 1919 (when the first such treaty was signed) and 1972 (when UNESCO's Convention on Cultural Heritage marked the global consolidation of the culture concept). The uses of a concept as widespread as "culture" can of course be charted through many other textual sources. But as a means of interrogating the uses of the culture concept in state-to-state relations these treaties are a uniquely valuable source. The treaties used concepts like civilization and culture in their carefully negotiated wording. As binding agreements under international law, such treaties regulated cross-border cultural flows and forged networks of exchange and obligation. These treaties are an interesting source, moreover, precisely because these texts were produced by diplomats, rather than cultural producers or "intellectuals" in a traditional sense. Examining cultural treaties allows us to study the moment when diplomats and statesmen from two (or more) countries agreed on the nature and content of the "culture" they planned to exchange. That moment of negotiation was also, I submit, an important moment for the transnational, trans-linguistic articulation of concepts.

In this project, I study these treaties with the large-scale, quantitative methods of the digital humanities, as well as with the tools of textual and conceptual analysis associated with the study of intellectual history. The project seeks, in other words, to conduct what Franco Moretti has called a "distant reading" of the treaties, as well as a close reading of a selection of these documents (and of archival material related to their preparation). In my paper for DH Nordic 2018, I will outline the topic, goals, and methods of the project, focusing on the ways we (that is, my colleagues at Umeå University's HUMlab and I) are approaching the "distant reading" part of this study of global intellectual history.

We explore the source material offered by these treaties by approaching it as two distinct data sets. First, we conduct quantitative analysis of the basic information, or "metadata" (countries, date, topic, etc.) from the complete set of treaties on cultural matters between 1919 and 1972, approximately 300 documents. Our source for this information is the World Treaty Index (www.worldtreatyindex.com). We ask several questions of this data. Which countries signed the most cultural treaties? At what historical moments were cultural treaties more or less frequently used? How does

the quantity of cultural treaties compare with that of other treaty types, like trade treaties, or treaties regulating (only) educational or scientific exchange? This data can also help identify the emergence of networks, or in some cases webs, of bilateral cultural treaties. Visualizing these networks (using tools such as Cytoscape or Gephi) will allow me to pose interesting questions by comparing them to any number of other transnational systems. How, for example, does the map of cultural agreements compare to that of trade treaties, military alliances, or to the transnational flows of cultural goods, capital, or migrants?

Second, to chart the development of concepts, we will observe the changing use of key terms through quantitative analysis of the treaty texts. The treaty texts (digital versions of most which can be found online) will be limited to four subsets: a) Britain and France, 1919-1972; b) India, 1947-1972; c) the German Reich (1919-1945) and the two German successor states (1949-1972); and d) UNESCO's multilateral Conventions (1945-1972). This selection is designed to approach a global perspective while taking into account practical factors, such as language and accessibility. By treating a large group of cultural treaties as several distinct text corpora and, perhaps, as a single text corpus, we will be able explore the treaties using several textometric methods, including measuring word frequencies, identifying co-occurence, and keyword extraction. This quantitative analysis should allow us to map the use of important concepts and phrases over time. Is it the case that “civilization” drops out of these documents, to be replaced by “culture”? Likewise, measuring the similarity of treaties to one another may enable to isolate which treaties acted as models that were later copied by other nations. Since the creators of each text are countries (and thus physical places) we can chart the changing word usages geographically. Using named-entity recognition might allow us to further link the texts to maps through geo-parsing. Spatializing the data allows us to ask which countries pioneered the transition to “culture” in international relations. In relation to which other countries did they do this? Can we identify particular groupings of countries (by continent, or by political ideology) that used the culture concept in similar ways?

Finally, we will seek to identify themes in the treaty texts through topic modeling. Textual data from these sources is of high enough quality for automatic part-of-speech tagging, enabling elimination of non-essential words as well as stemming (grouping together various forms of a word that share the same root). These preparations will enable us to run a more powerful, targeted form of text analysis through natural language processing tools like MALLET (mallet.cs.umass.edu). Over all, our use of text analysis seeks (a) to offer insight into the changing usage and meanings of concepts like “culture” and “civilization” in international documents; (b) to identify which areas of cultural activity were regulated by the treaties over time and by world region; and (c) to clarify whether “culture” was used in a broad, anthropological sense, or in a narrower sense to refer to the realm of arts, music, and literature. This aspect of the project raises interesting challenges, for example regarding how best to manipulate a multi-lingual text corpus (with texts in English, French, and German, at least).

In these ways, the project seeks to contribute to our understanding of how the concept of culture that guides today's international society developed. It also explores how digital tools can help us ask (and eventually answer) questions in the field of global intellectual history.

Finnish Aesthetics in Academic Databases

Darius Pacauskas & Ossi Naukkarinen

ABSTRACT

The major academic databases such as Web of Science and Scopus are dominated by publications written in English, often by scholars affiliated to American and British universities. As such databases are repeatedly used as basis for assessing and analyzing activities and impact of universities and even individual scholars, there is a risk that everything published in other, especially minor languages, will be sidetracked. Standard data-mining procedures do not notice them. Yet, especially in humanities, other languages and cultures have an important role and scholars publish in various languages.

The aim of this research project is to critically look into how Finnish aesthetics is represented in scientific databases. What kind of picture of Finnish aesthetics can we draw if we rely on the metadata from commonly used databases?

We will address this general issue through comparing metadata from overall five different databases, in two different languages, English and Finnish, and form a picture of several different interpretations of an academic field, aesthetics - or "estetiikka" in Finnish. To achieve this target we will employ citation analysis, as well as text summarization techniques, in order to understand the differences between the largest scientific databases in the world and the largest Finnish ones. We will employ data from Scopus, Google Scholar, Elektra, Arto, Helka, and some internal university databases - Tuhat, Acris, and CRIS. But most of the emphasis will be put on Scopus and Elektra databases that will be used in several parts of analysis. These databases were chosen for an extensive analysis due to availability of needed metadata (references) and because they belong among the largest in their own areas: Scopus - international, Elektra - Finnish. Additionally Scopus has convincing API for gathering data.

Moreover, we chose and will present some of the most influential Finnish aestheticians and analyze their publications record in order to understand to what extent the scientific databases can represent Finnish aesthetics. These Finnish aestheticians were chosen by the second author of this paper, Professor Ossi Naukkarinen, based on his extensive experience and knowledge of the field.

We will present:

- A world map based on Scopus that shows how scholars from various countries cite other scholars from other countries. The map is based on the most cited references as indexed in Scopus. It shows how geographically and culturally biased Scopus is as regards fields such as aesthetics. For example, non-English references are practically missing.
- A selection of Finnish aestheticians and the distribution of their works across different databases. This comparison shows how different pictures different databases offer and hence none of them can be trusted as a single source of information.
- A social network map that is created by bibliographic coupling of data from Scopus and Elektra as well as from open sources that include estheticians' works, for example, publications in journals where Finnish aestheticians tend to publish, such as the Finnish journal Synteesi and the US-based e-journal Contemporary Aesthetics. The map is based on references and text summarization.

Through this, we will present 1) two different maps containing actors and works recognized in the field, and 2) an overview of the main Finnish aestheticians and their works indexed in different databases.

For these goals, we will collect metadata from both Scopus and Elektra databases with references from each relevant article. Relevant articles will be located by using keyword “aeshetics” or the Finnish equivalent “estetiikka”, as well as identifying scientific journals focusing on aesthetics. We will perform citation analysis to explore in which countries which publications are cited, based on Scopus data. This comparison will allow us to understand what are the most prominent works for different countries, as well as to find the countries in which those works are developed, e.g., works that are acknowledged by Finnish aestheticians according to international databases.

Later, we will perform a citation analysis with the data gathered from the Finnish scientific database Elektra. Results will allow us to compare most cited works from both databases, as well similarity of references used. Moreover, it will indicate distribution between the cited Anglo-American texts and the ones written in Finland and/or in Finnish. Thus we could understand which language-family sources Finnish aestheticians rely on in their works, and what works in particular form the basis of their work. Further, we will apply text summary techniques to see the differences in the topics both databases are discussing. Here, we will face the question which databases are good sources for these kinds of analyses, if any? Do we have reliable databases and if we have, who has access to them? Neither Scopus, nor Elektra, for example, seem to cover everything that is relevant.

Furthermore, we will present a list of names of some of the most influential Finnish aestheticians, and their works (as provided by the databases). We will perform searches within five databases to understand how much of their works are covered.

As an additional contribution we developed an interactive web-based tool, which is accessible on http://dhoa.aalto.fi/dhn/ to represent results of this research. Such tool will give an opportunity for aesthetics researchers to explore Finnish aesthetics field through our established lenses and also comment on possible gaps in the pictures offered by the databases. It is possible that databases only give a very partial picture of the field and in this case new tools should be developed in co-operation with researchers. The similar situation might be true also in other sub-fields of humanities where activities in other languages than English are usual.

Abstract, DHN 2018 Helsinki
Conference track: Short paper
Theme: History

Roger Mähler, System developer Humlab
Umeå University roger.mahler@umu.se

Fredrik Norén, PhD candidate
Department of Culture and Media / Humlab Umeå
university fredrik.noren@umu.se@umu.se

The World According to the Popes: A Geographical Study of the Papal Documents, 2005–2017

This is a method-oriented project that seeks to explore what an atlas of the popes would look like. How can geography and the Vatican be studied together through texts? Can one study places in texts to map latent meanings of political and religious ambitions, and anticipate evolving trends? Is spatial analysis a way to better understand a closed institution such as the Holy See?

The Vatican is often associated with conservative stability. The papacy has, after all, managed to prevail while states and supranational organizations have come and gone. At the same time, the Vatican has shown remarkable capacity to gradually adapt to scientific paradigms as well as a changing world. This complexity also reflects geopolitical strategies of the Catholic Church. During the twentieth century, for example, the church state has expanded its global presence. When John Paul II was elected pope in 1978, the Vatican City had full diplomatic ties with 85 states. In 2005, when Benedict XVI was elected, that number had increased to 176. Moreover, the papacy has now formal diplomatic relations with the European Union, and is represented as a permanent observer to various global organizations including United Nations, the African Union, the World Trade Organization, and has even obtained a special membership in the Arabic League (Agnew, 2010; Barbato, 2012). In fact, the emergence of an international public sphere has been utilized by the Holy See, and significantly increased its soft power (Barbato, 2012).

As the geopolitical conditions and ambitions of the Vatican City are changing, what happens with its perception of regions and places? Does it constitute fixed historical patterns, or is it evolving, and changing when a new pope is elected? Inspired by Franco Moretti, this study departs from the notion that making connections between places and texts "will allow us to see some significant relationships that have so far escaped us" (Moretti, 1998: 3). The basis of the analysis is all English translated papal documents from Benedict XVI (2005–2013) and Francis (2013–), retrieved from the Vatican webpage (http://www.vatican.va/holy_father/index.htm).

Methodological Preparations: Scraping Data and Extracting Entities

From a technical point of view, the empirical material used in this study has been prepared in three steps. First, all web page documents in English have been downloaded, and the (proper) text in each document has been extracted and stored. Second, the places mentioned in each text document have been identified and extracted using the Stanford Named Entity Recognizer (NER) software. Third, the resulting list of places has been manually reduced by merging the most obvious name variations of the same place (e.g. "Sweden" and "Kingdom of Sweden").

The Vatican's communication strategies differ from, let's say, those of the daily press or the parliamentary parties, in the sense that they have a thousand-year perspective, or work from the point of view of eternity (Hägg, 2007). This is reflected on the Vatican's webpage, which is immensely informative. Text material from all popes since the late nineteenth century are publicly accessible online, ranging from letters, speeches, bulls to encyclicals, and all with a high optical character recognition (OCR) quality. Since the Holy See always has been a, according Göran Hägg, "mediated one man show", it makes sense to focus on a corpus of texts written or spoken by the popes in order to study the Vatican's notion of, basically, everything (Hägg, 2007: 239). The period 2005 to 2016 is pragmatically chosen because of its comprehensive volume of English translated papal documents.

Before this period, as Illustration 1 shows, you basically need to master Latin or Italian. While, for example, the English texts from John Paul II (1978–2005) equals to two million words, the corpus of Benedict XVI (2005–2013) together with current pope Francis sum up to near 59 million words, spread over some 5000 documents.

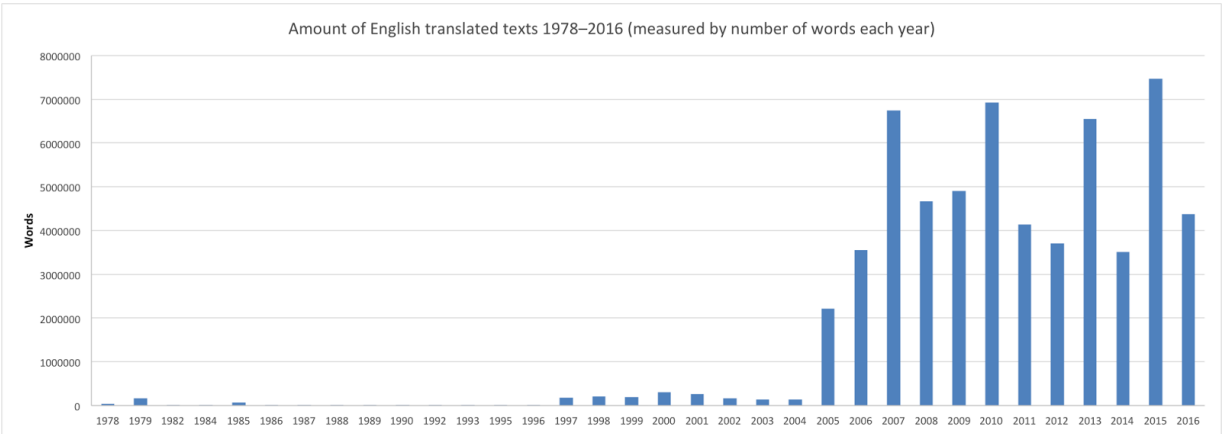


Illustration 1. The table shows the change in English translated text material available at the Vatican webpage.

The text documents were extracted, or "scraped", from the Vatican web site using scripts written in the Python programming language. The *Scrapy* library was used to "crawl" the web site, that is, to follow links of interest, starting from each Pope's home page, and download each web page that contains a document in English. The site traversal (crawling) was governed by a set of rules

specifying what links to follow and what target web pages (documents) to download. The links (to follow) included all links in the left side navigation menu on the Pope's home page, and the "paging" links in each referenced page. These links were easily identified using commonalities in the link

URL's, and the web pages with the target text documents (in HTML) were likewise identified by links matching the pattern ".../content/*name-of-pope*/en/.../documents/". The *BeautifulSoap* Python library was finally used to extract and cleanse the actual text from the downloaded web pages. (The text was easily identified by a '.documento" CSS class.) In the next step we ran the Stanford Named Entity Recognizer on the collected text material. This software is developed by the Stanford Natural Language Processing Group, and is regarded as one of the most robust implementation of *named entity recognition*. That is, the task of finding, classifying and extracting (or labeling) "entities" within a text. Stanford NER uses a statistical modeling method (Conditional Random Fields, CRFs), has multiple language support, and includes several pre-trained classifier models (new models can also be trained). This study used one of the pre-trained models, the 3 class model (location, person and organization) trained on data from CoNLL 2003 (Reuters Corpus), MUC 6 and MUC 7 (newswire), ACE (newswire, broadcast news), OntoNotes (various sources including newswire and broadcast news) and Wikipedia. This is the reason why, for example, "Hell" was not identified as a place, or why "God" rarely was a person, or a place. However, since the presentation focuses on what could be labeled as "earthly geography", this was not considered a problem for the analysis. Stanford NER tags each identified entity in the input text with the corresponding tag (location, person and organization). These tagged entities were then extracted from the entire text corpus and stored in a single spreadsheet file, aggregated on the number of occurrences per entity and document. (The stored columns were document name, document year, type of document, name of pope, entity, entity classifier, and number of occurrences.)

Even though some of the places identified by Stanford NER were difficult to assess whether they were in fact persons or organizations, they were still kept for the analysis. Furthermore, abstract geographical entities such as "East", or very specific ones (but still difficult to geographically identify) like "Beautiful Gate of the Temple", or an entity like "Rome-Byzantium-Moscow", which could be interpreted as a historic political alliance; all these places were kept for the analysis. After all, in this study the interest lies in the general connections between places, not the rare ones, which easily disappear in the larger patterns.

Papa Analytics

The question of how to deal with "Europe" has always constituted an issue for the Vatican. For centuries, the Catholic Church was essentially a European church, at least until Columbus' "discovered" America. The Holy See has always considered Europe as a whole, and struggled to deal with, for example, the rise of nation states during the nineteenth century, and the later division

between east and west Europe during the twentieth century. The Vatican perceive Europe as deeply rooted in ”her Christian soul”, and in the conviction that Christianity shaped European civilization, culture and values, which the Holy See considers to be universal (O’Mahony, 2009). Hence, this presentation, which constitute a work in progress, will use Europe as a case and focuses on the papal notion’s of the continent, and how Europe’s geographical associations and connections to various places have been expressed and evolved from Benedict XVI to Francis. Based on the preparations discussed above the analysis consists of three parts, with different methodological approaches, which utilizes the identified place entities and the entire text corpus. The aim is to explore if these methods could complement the scholarly work concerning the geopolitics of the Vatican.

First, the study introduces some spatial reflections of the recent papacy using simpler methods to trace, for example, how frequencies of certain places change in the texts, and which contexts (document types) are most geographical oriented. Furthermore, how the geographical density has changed over time, that is, how many places (total or unique ones) are mentioned per documents or per 1000 words.

Second, the analysis will utilize clusters of “co-occurring” places, based on places mentioned in the same document, to study (Gephi based) networks of center and periphery. Since most individual papal texts are dedicated to a certain topic, one can assume that places in a document have something in common. The term frequency-inverse document frequency (tf-idf) weighting is used as a measure of how important a place is in a specific document, and used in the co-occurrence computation. This unfolds latent geographical networks, as it is articulated by the papacy, with centers and peripheries, and both sacred and geopolitical aspects.

Last, the presentation will explore differences in linguistic associations with certain places, as they are expressed through Benedict XVI and pope Francis. For this purpose, the study utilizes word2vec, which is a method developed by a team at Google in 2013, to produce word embeddings (Mikolov et al, 2013). Simply put, the algorithm positions the vocabulary of a corpus in a high-dimensional vector space based on the assumption that “words which are similar in meaning occur in similar contexts” (Rubenstein & Goodenough, 1965: 627). This enables the use of basic numerical methods to compute word (dis-)similarities, to find clusters of similar words, or to create scales on how (subsets of) words are related to certain dichotomies. This study investigates different word associations regarding, for example, “Europe” and “Africa” (x axis), and “Public” and “Private” (y axis).

References

Agnew, J. (2010). Deus Vult: The Geopolitics of the Catholic Church. *Geopolitics*, *15*(1), 39–61.

Barbato, M. (2012). Papal Diplomacy : The Holy See in World Politics. *IPSA XXII World Conference of Political Science*, (2003), 1–29.

Finkel, J.R. Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

Florian, R., Ittycheriah, A., Jing, H. and Zhang, T. (2003) Named Entity Recognition through Classifier Combination. *Proceedings of CoNLL-2003*. Edmonton, Canada.

Hägg, G. (2007). *Påvarna : två tusen år av makt och helighet*. Stockholm: Wahlström & Widstrand.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, 1–12.

Moretti, F. (1998). *Atlas of the european novel: 1800–1900*. New York: Verso.

O'Mahony, A. (2009). The Vatican and Europe: Political Theology and Ecclesiology in Papal Statements from Pius XII to Benedict XVI. *International Journal for the Study of the Christian Church*, *9*(3), 177–194.

Rodriquez, K. J., Bryant, M., Blanke, T., & Luszczynska, M. (2012). Comparison of Named Entity Recognition tools for raw OCR text. *Proceedings of KONVENS 2012 (LThist 2012 Workshop)*, *2012*, 410–414.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

EXPLORING COUNTRY IMAGES IN SCHOOL BOOKS: A COMPARATIVE COMPUTATIONAL ANALYSIS OF GERMAN SCHOOL BOOKS IN THE 20TH CENTURY

Dr Kimmo Elo, University of Helsinki, Finland/ Political History
M.A. Virpi Kivioja, University of Turku/ Contemporary History

Abstract for the 3rd Digital Humanities in the Nordic Countries Conference
Helsinki, Finland, 7-9 March, 2018

Abstract

This paper is based on an ongoing PhD project entitled "An international triangle drama?", which studies the depictions of West Germany and East Germany in Finnish, and depictions of Finland in West German and East German geography school books in the Cold War era. The primary source material consists of Finnish, German (both West, East and unified) geography school books published between 1946 and 1999.

Contrary to traditional methods of close reading thus far applied in school book analysis, this paper presents an exploratory approach based on computational analysis of a large book corpus. The corpus used in this presentation consists of 91 school books in geography used in the Federal Republic of Germany between 1946 and 1999 (n=77), and in the German Democratic Republic between 1946 and 1990 (n=14). The corpus has been created by digitising all books by applying OCR technologies on the scanned page images. The corpus has also been post-processed to correct OCR errors and to add metadata. The corpus has ca. 540,000 tokens and ca. 16,000 unique lemma.

The main aim of the paper is to extract and analyse conceptual geocollocations. Such an analysis focuses on how concepts are embedded geospatially on the one hand, how geographical entities (cities, regions, etc.) are conceptually embedded, on the other. Regarding the former, the main aim is to examine and explain the geospatial distribution of terms and concepts. Regarding the latter, the main focus is on the analysis of concept collocations surrounding geographical entities.

This paper exploits digital methods in three different domains. First, we use standard methods of text mining to extract geographical clusters, i.e. text blocks with geographical concepts (names of different regions, cities etc.) in the centre. Second, we create geospatial data by enriching non-geographical concepts in each cluster with geocodings. And third, we use statistical and digital methods to explore geography-based differences in the use of other concepts and to visualise our main results.

Since the paper is based on an early-stage research project, it will present preliminary results only. We hope to be able to evidence correlations between regions/places and the concepts used to provide information about the region. We also seek to visualise differences in geospatial distribution of core concepts used in the school books.

Concerning historical research from a more general perspective, one of the main objectives of this paper is to exemplify and discuss how computational methods could be applied to tackle research questions typical for social sciences and historical research. The paper is motivated by the big challenge to move away from computational history guided and limited by tools and methods of computational sciences toward an understanding that computational history requires computational tools developed to find answers to questions typical and crucial for historical research.

Keywords: Exploratory analysis, Computational history, School book analysis, Geospatial modelling

Games, particularly massively multiplayer online role-playing games (MMORPGs), are about community; the micro interactions and collective experiences that occur between members of the community within the space of the game, and sometimes beyond. These experiences and interactions contribute to the reciprocal building of both the player's reputation in a particular game and the reputation of the game among the game's community. My work uses *Lord of the Rings Online* (*LOTRO*), a digital game adaptation of J.R.R. Tolkien's Middle-earth, to explore how reputation and community impact the gamespace. Regrettably, scholars working at the intersections of Tolkien Studies, adaptation theory, and game studies have not yet attended to the importance of reputation and community in this context. The translation of the narrative across mediums, the transmedial expansion of the narrative, and the reputation of the end product as an adaptation are often the focus of study. And while several scholars note the reputation of *LOTRO* as an adaptation of Tolkien's textual universe, less recognized is the *impact* of this change in space on the participant's immersion in and the community's engagement with the world of Middle-earth. In an attempt to address this gap, this paper considers how the spatial structures of reputation - or the components that shape reputation - impact engagement in and with the play environment.

There are two types of reputation addressed in this paper: player reputation, which includes affordances like the player's avatar, their information bubble, and their fellowships; and game reputation, which includes elements like the game's interface, graphics, and adaptation of space. How do the spatial structures of reputation impact engagement *in* the play environment? And why do they impact engagement *with* the gamespace? The "how" merits a closer analysis of the player's reputation by examining how the game's social architectures - or the mechanisms that afford in-game communication - affect the player: the effect of interface on social relations

and on community formation in the game influences social interaction between players. The "why" requires a consideration of the game's reputation: the degree to which *LOTRO* convinces its players that it is Tolkien's Middle-earth has a direct impact on two factors: first, the success of seamless player integration into this virtual world. And second, the formation of a digitally mediated space that functions as a place of social gathering.

The results of my work are twofold: first, the interlocutionary accoutrements that comprise player reputation attest to the ethos of individual players. These interlocutionary accoutrements impact engagement between players, acting as the channel through which players exchange and communicate information with each other. Second, the interface of the game grants the player access to an architected space. The components of the game's reputation imbricate to establish this space as Middle-earth, and encourage the community to engage with it as such. In short, my work illustrates that the spatial structures of reputation immerse the player into the gamespace, and shape it around a sense of social cohesion grounded in player-to-player micro interactions and community-wide collective experiences.

Its your data, but my algorithms

ABSTRACT

The world is increasingly digital, but the understanding of how the digital affects everyday life is still often confused. Digitalisation is sometimes optimistically thought as a rescue from hardships, be it economical or even educational. On the other hand, digitalization is seen negatively as something one just can't avoid. Digital technologies have replaced many previous tools used in work as well as in leisure. Furthermore, digital technologies present an agency of their own into the human processes as marked by David Berry. Through manipulating data through algorithms and communicating not only with humans, but other devices as well, digital technology presents new kind of challenges for the society and individual. These digital systems and data flows get their instructions from the code that runs on these systems. The underneath code itself is not objective nor value-free and carries own biases as well as programmers, software companies or larger cultural viewpoints objectives. As such, digital technology affects to the ways, we structure and comprehend, or are even able to comprehend the world around us.

This article looks at the surrounding digitality through an artistic research project. Through using code not as a functional tool but in a postmodern way as a material for expression, the research focuses on how code as art can express the digital condition that might otherwise be difficult to put into words or comprehend in everyday life. The art project consists of a drawing robot controlled by EEG-headband that the visitor can wear. The headband allows the visitor to control the robot through the EEG-readings read by the headband.As such the visitor might get a feeling of being able to control the robot, but at the same time the robot interprets the data through its algorithms and thus controls the visitor's data and the robot's actions.

The aim of this research project is not to focus on the interaction design or illustrating some aspects of brain science, instad the project aims to give perspectives to the everydayness of digitality. It wants to question how we comprehend digital in everyday life and asks how we should embody digitality in the future. The benefits of artistic research are in the way it can broaden the conceptions of how we know and as such can deepen one's understanding of the complexities of the world. Furthermore, artistic research can expand the meaning to alternative interpretations of the research subjects. As such, this research project aims at the same time to deepen the discussion of digitalization and to broaden it to alternative understandings. The alternative ways of seeing a phenomenon, like digitality, are essential in the ways future is developed.

Keywords: *creative coding, artistic research, digital humanities, postdigital, digitalisation*

Prose Rhythm in Narrative Fiction: the case of Karin Boye's *Kallocain*

Linguistic rhythm is an important feature of prose fiction, both at the level of narration and speech. The concept of 'linguistic rhythm' is notoriously vague, but one definition has been suggested (Holm, 2015), at least of one aspect: graphic rhythm in prose fiction can be regarded as the variation in syntactic structure of the sentence.

A good case in point to employ this definition is Karin Boye's (1900-1941) last novel *Kallocain* (1940), an icily dystopian depiction of a totalitarian future. The protagonist Leo Kall first embraces this system, but for various reasons rebels against it. The peripety comes when he gives a public speech, questioning the State. It has been pointed out (by the linguist Olof Gjerdman) that the novel – which is narrated in the first-person mode – from exactly this point on is characterized by a much freer rhythm (Gjerdman, 1942). This paper sets out to test this hypothesis, moving on from a discussion of the concept of rhythm in literary prose to an analysis of various indicators in different parts of *Kallocain.*

Work on this project is still in early stages. So far we have performed preliminary experiments with simple text quality indicators, including most of Holm's (2015) indicators, using measures like word, phrase, and sentence length, and the proportion of different types of punctuation marks. For all these indicators we have compared the first half of the novel, up until the speech, the second half of the novel, and as a contrast also the "censor's addendum", which is a short last chapter of the novel, written by an imaginary censor. Since the novel is narrated in the first-person mode, it can be expected that the rhythm change only affect the main narrative parts and the speech of the narrator, not the speech of the other characters, since they do not undergo the same change as the narrator. However, speech attribution is a difficult task, and a research field of its own (Elson & McKeown, 2010), which we have not attempted. Instead we opted to only compare the narrated parts of the novel, ignoring speech passages. It is always clearly marked where a speech passage starts, but not always where it ends. So when unclear, we removed all parts until the next paragraph, thus

making sure that all speech is removed, but possibly also removing small additional parts of the text. The need to separate speech and non-speech has also been pointed out by Holm (2015). We did note that removing speech made only a minor difference to the difference in indicators in the two parts, though.

For most of the indicators we find no or only minor differences between the two major parts of the novel. Only three indicators seem to point to a more strict rhythm in the first half. The proportion of long words, both as counted in characters and syllables, are considerably higher in the first half. For instance, the percentage of words with at least five syllables is 2.05% in the first half, and 1.03% in the second half. The sentence length is longer in the first half than the second, with 18.3 tokens per sentence compared to 16.2. The last part contains a proportionally higher rate of past perfect tense, counted as the number of instances of the word “hade” (“had”) used as an auxiliary, 8.9 versus 6.5 per 1000 words. The past perfect tense has been used as an indicator of less formal text (Holm 2015). The other indicators with a major difference do not support the hypothesis, however. In the first half, the sections are shorter, there are proportionally more speech utterances, and there is a higher proportion of three consecutive dots (...), which are often used to mark hesitation. If we compare these two halves to the censor's addendum, however, we can clearly see that the addendum is written in a stricter way, with for instance a considerably higher proportion of long words (4.90% of the words have more than five syllables) and more than double as long sentences.

In future analysis, we plan to use more fine-tuned indicators, based on a dependency parse of the text, from which we can explore issues like the proportion of main clauses and sub clauses and the number of arguments for different parts of speech. We also plan to explore the variation in our indicators, rather than just looking at averages, since this has been suggested in literature on rhythm in Swedish prose (Nordman, 2013).

Through this initial analysis we have also learned about some of the challenges of analyzing literature. For instance, it is not straightforward to separate speech

from non-speech, since the end of utterances are often not clearly marked in *Kallocain*, and free indirect speech is sometimes used. We think this would be important for future analysis, as well as attribution of speech (Elson & McKeown, 2010), since the speech of the different protagonists cannot be expected to vary in the two parts to the same degree. We also want to compare the rhythm in Kallocain to other novels by Boye. In further work, we want to be able to explore these indicators on a wider range of novels than only Boye, allowing large scale investigations on rhythm in prose, which have up until now mostly been explored on a smaller scale (Nordman, 2013; Holm, 2015).

References

Elson, David K. and McKeown, Kathleen R. (2010) Automatic Attribution of Quoted Speech in Literary Narrative. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*. The AAAI Press, Menlo Park, pp 1013–1019.

Gjerdman, Olof (1942) Rytm och röst. In *Karin Boye. Minnen och studier*. Ed. by M. Abenius and O. Lagercrantz. Stockholm: Bonniers, pp 143–160.

Holm, Lisa (2015) Rytm i romanprosa. In *Det skönlitterära språket: Tolv texter om stil*. Ed. by C. Östman. Stockholm: Morfem, pp 215–235.

Nordman, Marianne (2013) *Allting är en spegel. Jarl Hemmers poesi och prosa i språkligt återsken*. Skrifter utgivna av Svensk-Österbottniska samfundet, nr 71.

Authors: Carin Östman, Sara Stymne, Johan Svedjedal (Uppsala University)

Digital Humanities in the Nordic Countries

3rd Conference

http://heldig.fi/dhn-2018/

Short paper Abstract: Digital archives and the learning processes of performance art

Tero Nauha

In this presentation, the process of learning performance art is articulated in the contextual change that digital archives have caused starting from the early 1990s. It is part of my postdoctoral research, artistic research on the conjunctions between divergent gestures of thought and performance, done in a research project *How to Do Things with Performance?* funded by the Academy of Finland.

Since performance art is a form of 'live art', it would be easy to regard that the learning processes are also mostly based on the physical practice and repetition. However, the origins of contemporary performance art are closely connected with the 1960's and 70's conceptual art and video-art. Therefore, the pedagogy of performance art has been tightly connected with the development of media from the collective use of the Portapak video cameras and to use the digital archive methods by such figures like William Forsythe. On the use of archive such digital journals of artistic research like Ruukku-journal or Journal for Artistic Research create a different respect to field, also.

In this short documentation I will focus through few examples on the pedagogy of performance art. The vast amount of historical archive materials has become accessible to students through internet, no matter where their physical location might be. My point of view is based on practical notions that I have gathered from the teaching of performance art from BA to PhD level research projects. Through mediatized learning processes, makes the archival material generic, easy access and possible for speculation what performance art is. However, this learning processes makes the artistic practices actualization of the virtual, but limited material. The digitalization is a modality of this process, where the necessity of embodied practice is still crucial. This learning process of performance art is not done through resemblance and imitation but doing *with* someone or something else and then developed in response or in diffraction with the digital archive materials.

Tero Nauha
teronauha@gmail.com
Doctor of Arts and a postdoctoral fellow at the Academy of Finland funded postdoctoral research project 'How To Do Things With Performance'
Theatre Academy of the University of the Arts Helsinki
www.teronauha.com
www.howtodothingswithperformance.wordpress.com

Poster Presentation Abstract
Digital Humanities in the Nordic Countries 2018

Katarina Pihlflyckt
Svenska litteratursällskapet i Finland
Zacharias Topelius Skrifter
katarina.pihlflyckt@sls.fi

ZACHARIAS TOPELIUS SKRIFTER – APPROACHING A DIGITAL SCHOLARLY EDITION THROUGH METADATA

This poster presents an overview of the database structure in the digital critical edition of Zacharias Topelius Skrifter (ZTS). The poster specifies what kind of information is available for the user and shows how the entity relations open a possibility for the user to approach the edition from other angles than the texts, by taking advantage of the information and descriptive metadata in indexing systems. Through this data, a historian can capture for example events, meetings between people or editions of books, as they are presented in Zacharias Topelius' (1818–1898) texts. Presented here are both already available features and features in progress.

ZTS comprises eight digital volumes hitherto, the first published in 2010. This includes the equivalent of about 8 500 pages of text by Topelius, 600 pages of introduction by editors and 13 000 annotations. The published volumes cover poetry, short stories, correspondences, children's textbooks, historical-geographical works and university lectures on history and geography. Genres still to be published include children's books, novels, journalism, diaries, religious and academic texts. It is freely accessible at topelius.sls.fi.

DATABASE STRUCTURE

The ZTS database structure consists of six connected databases: people, places, bibliography, manuscripts, letters and a chronology. So far, the people database consists of about 10 000 unique persons, and a possibility to link them to a family or group level (250 records). It has separate chapters for mythological persons (500 records) and fictive characters (250 records). The geographic database has 6 000 registered places. The bibliographic database has 6 000 editions divided on 3 500 different works, and the manuscript database has 1 400 texts by Topelius on 350 physical manuscripts. The letter database has 4 000 registered letters to and from Topelius, divided on 2 000 correspondences. The chronology of Topelius life has 7 000 marked events. The indexing of objects started in 2005, using the FileMaker system. The preliminary idea at that time was to give the users similar information as in indexes and annotations in printed scholarly editions of the same type. The work with finding more possibilities on how to use, link and present the data is in constant progress. Restrictions are mostly due to time-limiting factors within the editorial work. Challenges lay in the requirement of having shared data for different genres and in working coherently with a large amount of material for many years in the ever-developing digital world.

The bibliographic database is the most complex database, and follows the *Functional Requirements for Bibliographic Records* (FRBR) model. This means we are making a difference between the abstract work and the published manifestations (editions) of that work. The FRBR focuses on the content relationship and continuum between the levels; anything regarded a separate work starts as a new abstract record, from where its own editions are created. Within ZTS, the abstract level has a practical significance, in cases when it is impossible to determine to which exact edition Topelius is referring. Also taken in consideration is that for example articles and short stories can have their own independent editions as well as being included in editions (e.g. magazines, anthologies). This requires two different manifestation levels subordinated the abstract level; the regular editions and the texts included in other editions, the records of the latter type must always link to records of the former.

The manuscript database has a content relationship to the bibliographic database through the abstract entity of a work. A manuscript text can be regarded as an independent edition of a work in this context (a manuscript that was never published can easily have a future edition added in the bibliographic database). The manuscript text itself might share physical paper with another manuscript text. Therefore, the description of the physical manuscript is created on a separate level in the manuscript database, to which the manuscript text is connected. The display of the manuscript database is still under planning.

The letter database follows the FRBR example; an upper level presents the complete abstract correspondence between Topelius and another person, and a subordinated level describes each physical letter within the correspondence. It is also possible to attach additional corresponding persons to occasional letters. The letter database is currently only available through the manuscript descriptions for the separate published letters.

The people database connects to the letter database and the bibliographic database, creating a one-to-many relationship. Any writer or author has to be in the people database in order to have their information inserted into the other two databases. Within the people database there is also a family or group level, where the records of individual family members can be grouped, but in contrary to the letter database, this is not a superordinate level. The people database is available through the indexes and the annotated people in the XML files. The group level is still under construction and currently not included in the display view.

The geographic database follows a one-level structure. Places in the letter and manuscript databases are linked from the geographic database, in a similar way as people. The geographic database is available through the indexes and the annotated places in the text.

The chronology database contains manually added key events from Topelius' life, as well as short diary entries from various calendars during his life. The main content, however, is automatically gathered records from other databases, linked through the marked dates when Topelius works were published or when he wrote a letter or a manuscript. Dates of birth and death of family members and close friends can be linked from the people database. The chronology creates a timeline that would not only give the user key events from Topelius' life, but also links to the other database records. Encoded dates in the XML files (letters, diaries, lectures, manuscripts etc.) could lead the user directly to the relevant text passages. The work with the chronology is for the moment not prioritized and a timeline display currently not available.

POSSIBILITIES FOR THE USER

Approaching a digital scholarly edition with over 8 500 pages can be a heavy task, and many will likely use the edition more as an object to study, rather than texts to read. For a user not familiar with the content of the different volumes, but still looking for specific information, advanced searches and indexing systems offer an alternative path into the relevant text passages. The information in the ZTS database records helps provide a picture of Finland in the 19th century as it appears in Topelius' works and life. A future feature for users is freely access to the data in the published records through an API (Application Programming Interface). This will create opportunities for the user to take advantage of the data in almost any wanted way: to create a 19th century bookshelf, an app for the most popular 19th century names or a map of popular student hangouts in 1830's Helsinki.

Through the indexes formed by the linked data from the texts, the user can find all the occurrences of a person, a place or a book in the whole edition. One record can build a set of ontological relations, and the user can follow a theme, while moving between texts. A search for a person will provide the user with information not only about the person, his/hers close family and the possible personal relation to Topelius, but also where Topelius mentions this person, what he has to say about the person, or if they possibly meet or interact. Furthermore, the user will be able to see if this person was the author of a

book mentioned by Topelius in his texts, or if the editors at ZTS have used the book as a source for editorial comments. The user will also be able to find the person's possible correspondence with Topelius. The geographic index can help the user create a geographic ontology with an overview of Topelius' whereabouts through the annotated mentions of places in Topelius' diaries, letters and manuscripts. Map coordinates are not included in the database, partly because of time-limiting reasons and partly because of the difficulty in finding correct coordinates for places that have disappeared or changed since the 1800's. Available are instead descriptions of the places and the bigger entity where it is situated (e.g. city, region or country). A division is made between places in Finland and outside Finland. Uncertainty is encoded in the XML file in cases where the editor cannot be sure, whether a person, a place or a literary work that Topelius mentions is the same as linked to in the database, and the connection shown as uncertain in the view.

The entity relation between the bibliographic database and the manuscript database creates a complete bibliography over everything Topelius wrote, all known manuscripts and printed editions that relate to a specific work. So far, there are 900 registered independent works by Topelius in the bibliographic database; these works are implemented in 300 published editions (manifestations) and 2 900 text versions included in those manifestations or in other independent manifestations. The manuscript database consists of 1 400 manuscript texts, of which 600 are linked from 350 abstract works in the bibliography. The FRBR model offers different ways of structuring the layout of a bibliography according to the user's needs, either through the titles of the abstract works with subordinate manifestations, or directly through the separate manifestations. Topelius' texts on the abstract level are also marked according to genre (e.g. journalism, children's books etc.) and the bibliography can be arranged based on this. Furthermore, the bibliography points the user to the published texts and manuscripts of a specific work in the ZTS edition and to text passages where the author himself discusses the work in question. The bibliography will be available in 2019, together with a register over all works by other authors that Topelius is mentioning in his text, no matter whether we know the exact edition or only the abstract work.

The level of detail is high in the records. For example, we register different name forms and spellings (*Warschau* vs *Warszawa*). Such information is included in the index search function and thereby eliminates problems for the end user trying to find information. Topelius often uses many different forms and abbreviations, and performing even an advanced search in the texts would seldom give a comprehensive result in these cases. The letter database includes reference words describing the contents of the correspondences. Thus, the possibilities for searching in the material are expanded beyond the wordings of the original texts.

Poster Presentation Abstract
Digital Humanities in the Nordic Countries 2018

Pieter Claes, Per Stam, Elisa Veit
Svenska litteratursällskapet i Finland
Henry Parlands Skrifter
pieter.claes@sls.fi, per.stam@su.se, elisa.veit@sls.fi

Challenges in textual criticism and editorial transparency

Henry Parlands Skrifter (HPS) is a digital critical edition of the works and correspondence of the modernist author Henry Parland (1908–1930). The first part of the edition will be published in the autumn of 2018. The poster presents chosen strategies for communicating the results of the process of textual criticism in a digital environment. How can we make the foundations for editorial decisions transparent and easily accessible to a reader?

Textual criticism is by one of several definitions "the scientific study of a text with the intention of producing a reliable edition" (*Nationalencyklopedin*, "textkritik". Our translation.) When possible, the texts of the HPS edition are based on original prints whose publication was initiated by the author during his lifetime. However, rendering a reliable text largely requires a return to original manuscripts as only a fraction of Parland's works were published before the author's death at the age of 22 in 1930. Posthumous publications often lack reliability due to the editorial practices and sometimes primarily aesthetic solutions to text problems of later editors.

The main structure of the Parland digital edition is related to Zacharias Topelius Skrifter (topelius.sls.fi) and similar editions (e.g. grundtvigsværker.dk). However, the Parland edition has foregone the system of a – theoretically – unlimited amount of columns in favour of only two fields for text: a field for the reading text, which holds a central position on the webpage, and a smaller, optional, field containing, in different tabs, editorial commentary, facsimiles and transcriptions of manuscripts and original prints. The benefit of this approach is easier navigation. If a reader wishes to view several fields at once, they may do so by using several browser windows, which is explained in the user's guide.

The texts of the edition are transcribed in XML and encoded following TEI (Text Encoding Initiative) Guidelines P5. Manuscripts, or original prints, and edited reading texts are rendered in different files (see further below). All manuscripts and original prints used in the edition are presented as high-resolution facsimiles. The reader thus has access to the different versions of the text in full, as a complement to the editorial commentary.

Parland's manuscripts often contain several layers of changes (additions, deletions, substitutions): those made by the author himself during the initial process of writing or during a later revision, and those made by posthumous editors selecting and preparing manuscripts for publication. The editor is thus required to analyse the manuscripts in order to include only changes made by the author in the text of the edition. The posthumous changes are included in the transcriptions of the manuscripts and encoded using the same TEI elements as the author's changes with an addition of attributes indicating the other hand and pen (@hand and @medium). In the digital edition these changes, as well as other posthumous markings and notes, are displayed in a separate colour. A tooltip displays the identity of the other hand.

One of the benefits of this solution is transparency towards the reader through visualization of the editor's interpretation of all sections of the manuscript. The using of standard TEI elements and

attributes facilitate possible use of the XML-documents for purposes outside of the edition. For the Parland project, there were also practical benefits concerning technical solutions and workflow in using mark-up that had already, though to a somewhat smaller extent, been used by the Zacharias Topelius edition.

The downside to using the same elements for both authorial and posthumous changes is that the XML-file will not very easily lend itself to a visualization of the author's version. Although this surely would not be impossible with an appropriately designed stylesheet, we have deemed it more practical to keep manuscripts and edited reading texts in separate files. All posthumous intervention and associated mark-up are removed from the edited text, which has the added practical benefit of making the XML-document more easily readable to a human editor. However, the information value of the separate files is more limited than that of a single file would be.

The file with the edited text still contains the complete author's version, according to the critical analysis of the editor. Editorial changes to the author's text are grouped together with the original wording in the TEI-element choice and the changes are visualized in the digital edition. The changed section is highlighted and the original wording displayed in a tooltip. Thus, the combination of facsimile, transcription and edited text in the digital edition visualizes the editor's source(s), interpretation and changes to the text.

Sources

Nationalencyklopedin, "textkritik". http://www.ne.se/uppslagsverk/encyklopedi/lång/textkritik (accessed 2017-10-19).

Metadata Analysis and Text Reuse Detection: Reassessing public discourse in Finland through newspapers and journals 1771–1917

Poster at DHN 2018 conference

Presenters: Ginter, Filip (1); Kanner, Antti (2); Lahti, Leo (1); Marjanen, Jani (2); Mäkelä, Eetu (2); Nivala, Asko (1); Rantala, Heli (1); Salmi, Hannu (1); Sippola, Reetta (1); Tolonen, Mikko (2); Vaara, Ville (2); Vesanto, Aleksi (2)

Organisation(s): 1: University of Turku; 2: University of Helsinki

During the period 1771–1917 newspapers developed as a mass medium in the Grand Duchy of Finland. This happened within two different imperial configurations (Sweden until 1809 and Russia 1809–1917) and in two main languages – Swedish and Finnish. The *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (COMHIS) project studies the transformation of public discourse in Europe and in Finland via an innovative combination of original data, state-of-the-art quantitative methods that have not been previously applied in this context, and an open source collaboration model.

In this study the project combines the statistical analysis of newspaper metadata and the analysis of text reuse within the papers to trace the expansion of and exchange in Finnish newspapers published in the long nineteenth century. The analysis is based on the metadata and content of digitized Finnish newspapers published by the National library of Finland. The dataset includes full text of all newspapers and most periodicals published in Finland between 1771 and 1920. The analysis of metadata builds on data harmonization and enrichment by extracting information on columns, type sets, publications frequencies and circulation records from the full-text files or outside sources. Our analysis of text reuse is based on a modified version of the Basic Local Alignment Search Tool (BLAST) algorithm, which can detect similar sequences and was initially developed for fast alignment of biomolecular sequences, such as DNA chains. We have further modified the algorithm in order to identify text reuse patterns. BLAST is robust to deviations in the text content, and as such able to effectively circumvent errors or differences arising from optical character recognition (OCR).

By relating metadata on publication places, language, number of issues, number of words, size of papers, and publishers and comparing that to the existing scholarship on newspaper history and censorship, the study provides a more accurate bird's-eye view of newspaper publishing in Finland after 1771. By pinpointing key moments in the development of journalism the study suggest that while the discussions in the public were inherently bilingual, the technological and journalistic developments advanced at different speeds in Swedish and Finnish language forums. It further assesses the development of the press in comparison with book production and periodicals, pointing towards a specialization of newspapers as a medium in the period post 1860. Of special interest is that the growth and specialization of the newspaper medium was much indebted to the newspapers being established all over the country and thus becoming forums for local debates.

The existence of a medium encompassing the whole country was crucial to the birth of a national imaginary. Yet, the national public sphere was not without regional intellectual asymmetries. This study traces these asymmetries by analysing text reuse in the whole newspaper corpus. It shows which papers and which cities functioned as "senders" and "receivers" in the public discourse in this period. It is furthermore essential that newspapers and periodicals had several functions throughout the period, and the role of the public sphere cannot be taken for granted. The analysis of text reuse further paints a picture of virality in newspaper publishing that was indicative of modern journalistic practices but also reveals the rapidly expanding capacity of the press. These can be further contrasted to other items commonly associated with the birth of modern journalism such as publication frequency, page sizes and typesetting of the papers.

All algorithms and software will be made openly available online, and can be located through the project's repositories (https://comhis.github.io/ and https://github.com/avjves/textreuse-blast). The results of the text reuse detection carried out in BLAST are stored in a database that has already been opened at http://comhis.fi.

Legal issues regarding tradition archives: the Latvian case study

Līga Ābele, Anita Vaivade
Latvian Academy of Culture

A legally unchartered and potentially challenging territory is reached when the tradition archives, while implementing their digitation projects and because of the more and more varied nature of the collection items (text, audio, video, images, games, maps), find themselves at the crossroads between the computing and the cultural heritage, where the digital humanities are situated, i.e. by creating digital data bases of their collections and putting them on the Internet. It also happens when a tradition archive reaches out to the public to incite digital cooperation in encountering, transmitting and interpreting the collections. Because of the trans-border nature of the Internet, national and international legal framework can influence digitisation projects of the tradition archives and their possible interconnection.

Due to abundance and variety of materials, lack of sufficient resources to deal with these materials, possibilities provided by the Internet environment, as well as the general aim to involve more and more the public into its cultural heritage safeguarding, non-commercial crowdsourcing is increasingly present in the work of tradition archives and other cultural institutions, both in dealing with the existing data and making them accessible, as well as in encountering new testimonies and broadening the collections. Every time a social relationship is formed in a social environment, it can have a legal nature attached to it. It means, there is a possibility to qualify legally this relationship, in terms of its nature (its subject matter, mutual relationship between the parties and with the third persons as well as consequences). In crowdsourcing projects of tradition archives, the forming relationship could be devised in the following manner. The subjects (parties) of the relationship are archive institution and the members of the public. The subject matter of the relationship (a set of duties and obligations of the parties) may consist of entrusting of specific tasks to participants, such as transforming the content, describing the objects, synthetizing the knowledge and skills.

In order to achieve more clarity and overview of the crowdsourcing process, one of the ways to structure the legal statuses and relationships forming within the process, would be by vectors of the data streams (encountering/gathering/processing of data and its further transferring): roles of the parties – rights, obligations, mutual responsibility and responsibility towards third persons;

legal nature of the relationship. It would not only be interesting to take a closer look at the set of legal relationship forming within the process, if any, but also useful, for the project management to foresee the possible legal amplifications. It would be important to apprehend the legal nature of relationship between the organiser of the project and the participant, in order to understand legitimate mutual expectations and consequences, including in case of malfeasance. This regards the quality required for the result of the task, confidentiality, intellectual property issues etc. It may also regard consequences in relation to the third parties. It is useful to apprehend the legal environment also for the sake of knowledge about the default regulatory framework, any given project would fall within, in case no action is taken by the project management.

Crowdsourcing is not regulated as such in the Latvian law. However, there are several types of contractual relationship that involve one party entrusting other one with a specific task that *prima facie* might possess certain similarities with the relationships forming within a crowdsourcing project in the sphere of private law: *employment, voluntary work, assignment.* Taking a closer look at these contractual relationship permits to leave out at once the employment contract and the assignment contract for the following reasons. The very definition of the work involves the remuneration clause as a mandatory requirement for a relationship to be qualified as an employment contract (Latvian Labour Law, article 3 and 28, paragraph 1), as for the non-commercial crowdsourcing activities the voluntary participation is the key. The very reason the employment contract is mentioned here relates to it its possibly beneficial aspect regarding the copyrights and long-term crowdsourcing projects. According to the Article 12, paragraph 1, of the Latvian Copyright Law "*if an author has created a work performing his or her duties in an employment relationship, [...] the economic rights of the author may be transferred, in accordance with a contract, to the employer*." In contrast, outside the employment relationship there is a possibility to revoke these rights. In order to obtain the right to use a work, it is necessary for the user of the work to receive the permission of the right holder, issued both as a licensing agreement and as a licence. If a licensing agreement or a licence is not restricted as to time, the author or other right holder may terminate the licensing agreement or revoke the licence, giving a notice six months in advance. (Article 40, paragraph 1 and 2 and Article 44, paragraph 2 of the Latvian Copyright Law). Not considering this copyright aspect, taking into account the nature of a particular project or of data encountering, could prove to be a problematic aspect in future given the general long-term aim of the cultural heritage safeguarding which is to transfer the cultural heritage values to the future generations.

Because of its voluntary nature the contractual relationship formed between the organiser and the participant would also lack a basic element required in order for it to be qualified as the assignment contract (contract of work-performance) as stated in Article 2212 of the Civil Code of Latvia (one party undertakes, using its tools and equipment and *for a certain remuneration*, to perform for another party an order, the production of some product or the conducting to its completion of some activity.) Neither is the Civil Law helpful with its regulations of a gifting (donating), as according to its regulations a gift is a legal transaction whereby one person grants *valuable property* to another through generosity and without remuneration (Article 1912).

The closest legal description of a relationship forming in a crowdsourcing project of a tradition archive could be found in the Law on Voluntary Work applicable in Latvia since January 1st, 2016. According to its Article 2 the voluntary work is described as organised and voluntary physical or intellectual work performed in good will by a person without remuneration for the grater good of society without aim of gaining profit (unofficial translation). There are restrictions in its Article 3 on who can be the organisers: the NGO's, state and municipal institutions, political parties. According to this restriction the Latvian Archives of Folklore could not qualify for the status of organiser. There are also several restrictive aspects that need to be considered. Firstly, the voluntary work cannot replace work of an employee. Second, for a person between age 13 – 16 to legally participate in a project, there must be a written permission from this person's legal guardians. The law does not require a written agreement (with specific exceptions), the institution officially charged with implementing the law provides model agreement forms for short-term and long-term tasks.

Even though currently this law contains the closest legal definition to describe the relationship formed by the crowdsourcing activities between the organisers and participants, however in its text it does not consider and even contradicts specificity of the digital crowdsourcing projects, which would render problematic of not impossible its direct application. The obligations this law sets for organisers, including requirements on working environment safety etc. are not consistent with the very nature of the digital crowdsourcing. However, this law could be interesting for the tradition archive crowdsourcing projects to be used as a sort of a roadmap for framing the relationship between organisers and participants, as it provides formulations as to the beginning and end of a legal relationship, rights and obligations of the parties, including obligation to inform, confidentiality and quality requirements etc. It could be used in the agreement forms to be accepted before the person can register as participant to a crowdsourcing project.

It can be concluded that within current legal framework, there is no apparent legal instrument in Latvia to be applied directly (without interpreting) to the factual relationship arising between organiser and participant in a crowdsourcing project of tradition archives. Which means that specific care should be accorded by the project organisers form one hand – not to overcomplicate its implementations, on the other hand – to ensure that the basic aspects of the cooperation are established, notified and agreed. In case of tradition archives these framework provisions should also tackle two of the most important issues governed by the specific laws – copyright and data protection, as the activities of the parties are susceptible to fall within the scope of these regulations.

One of the main concerns while digitising the existing collections of tradition archives is the compliance of these databases with the regulatory framework of data protection. In case of Archives of Latvian Folklore, its institutional status is neither that of an archive, a museum or a library. Thus, there would be no specific public law provisions directly applicable to its activities in the field of personal data protection regulations reserved for these institutions. For instance article 12 of the Law on Archives titled “Accessibility and Use of Archival Records” is stated that a person has the right to request and obtain the information regarding other person's data subject, if *a written permit* has been received from that persons, as well as in cases specified by the Law. Also access is restricted for records containing sensitive personal data or other information on the private life of a person, if the use of personal data or information contained therein can significantly touch the private life of a person.

The Archive of Latvian Folklore is obliged to follow the general regulations of data protection. At the European Union level this field has undergone a legislative reform, hence starting May 25, 2018 a new General Data Protection Regulation will come into force (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC). As it is stated in Article 4 of the regulation, personal data means any information relating to an identified or identifiable natural person (data subject) and an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. In turn “processing” means any

operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

The terms of personal data protection a role of data holder is taken by the institution of tradition archives, and participants are users. Moreover, the very crowdsourcing project may fall within the scope of the Data protection regulations, in terms of the rights of its participants (i.e. right to erasure "right to be forgotten" according to Article 17 of the Regulation). Attention should be given to situations when natural person becomes identifiable while processing a specific item of a collection (for example, while transcribing etc.). Another case is when the content of a collection is co-created by the participant and the personal data of another natural persona might be processed without consent. Also specific rules apply through transmitting the databases to free access on the internet (for instance – how the participant is to insure that no data protection or copyright law is infringed in a material, submitted for collection). In this case – the participants should be very clearly notified about their obligations and possible consequences of breach, because the lack of accountability is identified as one of the major risks in crowdsourcing project.

The copyright law is the second issue arising from the factual relationship. In case of crowdsourcing projects of tradition archives, there are two main scenarios that could realise involving the copyright issues. In one case the item of the collection is object of copyright law, in another – work of the participant is object of copyright law. In creating the databases of tradition archives the organisers should take into account that especially in co-creative projects the work submitted by participants would most likely fall within the scope) of the protected works under copyright law (article 4 of the Copyright Law of the Republic of Latvia). The database itself would be considered as derived work and would be protected as such. In digitising data bases and transmitting them to public access, for the sake of continuity of collections the organisers should consider that among inalienable moral right of author there are right to the revocation of a work, that can be issued at any given time. There is specific (*sui generis*) regime in the law of protection of databases that regard the creation of databases of tradition archives, determining the rights of database developers and users. Also, the regulation of orphan works might be specifically interesting in case of tradition archive databases.

Use of a work of an author without the consent of the author and without remuneration is permitted among others in cases of a work is used for educational and research purposes and for the needs of libraries, archives and museums (Article 19). It is worth noting that only such works that have been published in Latvia and are not available commercially are permitted to be reproduced in a digital format, unless an agreement with the author determines otherwise. (Article 23, paragraph 1). To obtain the right to use a work, it is necessary to receive permission of the right holder in form of licensing agreement or a licence. The law determines requirements for the term, territorial scope and form of those agreements and licences. In case of crowdsourcing a thought should be given to the use of the Creative Commons copyright licenses that is a standardized tool to grant copyright permissions to a person's creative work with the emphasis of non-commercial use.

Even though discussion of the legal aspects of the crowdsourcing projects launched by the tradition archives still might seem to be too much of a theoretical kind of exercise, starting to apprehend the potential legal consequences just might support more confident managing of the future crowdsourcing projects in cultural heritage, i.e. tradition archives.

Heritage Here, K-Lab and intra-agency collaboration in Norway

Introduction

This paper aims to give an overview of an ongoing collaboration between four Norwegian government agencies, by outlining its history, its goals and achievements and its current status. In doing so, we will, hopefully, be able to arrive at some conclusions about the usefulness of the collaboration itself – and whether or not anything we have learned during the collaboration can be used as a model for – or an inspiration to – other projects within the cultural heritage sector or the broader humanities environment.

First phase – "Heritage Here" 2012 – 2015

Heritage Here (or "Kultur- og naturreise" as it is known in its native Norwegian) was a national project which ran between 2012 and 2015 (http://knreise.org/index.php/english/). The project had two main objectives:

1. To help increase access to and use of public information and local knowledge about culture and nature
2. To promote the use of better quality open data.

The aim being that anyone with a smartphone can gain instant access to relevant facts and stories about their local area wherever they might be in the country.

The project was a result of cross-agency cooperation between five agencies from 3 different ministries. Project partners included:

- the Norwegian Mapping Authority (Ministry of Local Government and Modernization).
- the Arts Council Norway and the National Archives (Ministry of Culture).
- the Directorate of Cultural Heritage and (until December 2014) the Norwegian Environment Agency (the Ministry of Climate and Environment).

Together, these partners made their own data digitally accessible; to be enriched, geo-tagged and disseminated in new ways. Content included information about animal and plant life, cultural heritage and historical events, and varied from factual data to personal stories. The content was collected into Norway's national digital infrastructure 'Norvegiana' (http://www.norvegiana.no/) and from there it can be used and developed by others through open and documented API's to create new services for business, tourism, or education. Parts of this content were also exported into the European aggregation service Europeana.eu (http://www.europeana.eu).

In 2012 and 2013 the main focus of the project was to facilitate further development of technical infrastructures - to help extract data from partner databases and other databases for mobile dissemination. However, the project also worked with local partners in three pilot areas:

- Bø and Sauherad rural municipalities in Telemark county
- The area surrounding Akerselva in Oslo
- The mountainous area of Dovre in Oppland county.

These pilots were crucial to the project, both as an arena to test the content from the various national datasets, but also as a testing ground for user community participation on a local and regional level. They have also been an opportunity to see Heritage Here's work in a larger context. The Telemark pilot was for example, used to test the cloud-based mapping tools developed in the

Best Practice Network “LoCloud” (http://www.locloud.eu/) which where coordinated by the National Archives of Norway.

In addition to the previously mentioned activities *Heritage Here* worked towards being a competence builder – organizing over 20 workshops on digital storytelling and geo-tagging of data, and numerous open seminars with topics ranging from open data and LOD, to IPR and copyright related issues. The project also organized Norway’s first heritage hackathon “#hack4no” in early 2014 (http://knreise.org/index.php/2014/02/27/hack4no-a-heritage-here-hackathon/). This first hackathon has since become an annual event – organized by one of the participating agencies (The Mapping authority) – and a great success story, with 50+ participants coming together to create new and innovative services by using open public data.

Drawing on the experiences the project had gathered, the project focused its final year on developing various web-based prototypes which use a map as the users starting point. These demonstrate a number of approaches for visualizing and accessing different types of cultural heritage information from various open data sets in different ways – such as content related to a particular area, route or subject. These prototypes are free and openly accessible as web-tools for anyone to use (http://knreise.no/demonstratorer/). The code to the prototypes has been made openly available so it can be used by others – either as it is, or as a starting point for something new.

Second phase – “K-Lab” 2016 – 2017

At the end of 2015 *Heritage Here* ended as a project. But the four remaining project partners decided to continue their digital cross-agency cooperation. So, in January 2016 a new joint initiative with the same core governmental partners was set up. Heritage here went from being a project to being a formalized collaboration between four government agencies. This new partnership is set up to focus on some key issues seen as crucial for further development of the results that came out of the *Heritage Here* project. Among these are:

- In cooperation develop, document and maintain robust, common and sustainable APIs for the partnerships data and content.
- Address and discuss the need for, and potential use of, different aggregation services for this field.
- Develop and maintain plans and services for a free and open flow of open and reusable data between and from the four partner organizations.
- In cooperation with other governmental bodies organize another heritage hackathon in October 2016 with the explicit focus on open data, sharing, reuse and new and other services for both the public and the cultural heritage management sector.
- As a partnership develop skillsets, networks, arenas and competence for the employees in the four partner organizations (and beyond) within this field of expertise.
- Continue developing and strengthening partnerships on a local, national and international level through the use of open workshops, training, conferences and seminars.
- Continue to work towards improving data quality and promoting the use of open data.

One key challenge at the end of the *Heritage Here* project was making the transition from being a project group to becoming a more permanent organizational entity – without losing key competence and experience. This was resolved by having each agency employing one person from the project each and assigning this person in a 50% position to the *K-Lab* collaboration. The remaining time was to be spent on other tasks for the agency. This helped ensure the following things:

- Continuity. The same project group could continue working, albeit organized in a slightly different manner.
- Transfer of knowledge. Competence built during Heritage here was transferred to organizational line of the agencies involved.
- Information exchange. By having one employee from each agency meeting on a regular basis information, ideas for common projects and solutions to common problems could easily be exchanged between the collaboration partners.

In addition to the allocation of human resources, each agency chipped in roughly EUR 20.000 as 'free funds'. The main reasoning behind this kind of approach was to allow the new entity a certain operational freedom and room for creativity – while at the same time tying it closer to the day-to-day running of the agencies.

Based on an evaluation of the results achieved in *Heritage Here*, the start of 2016 was spent planning the direction forward for *K-Lab*, and a plan was formulated – outlining activities covering several thematic areas:

Improving data quality and accessibility. Making data available to the public was one of the primary goals of the Heritage here project, and one most important outcomes of the project was the realisation that in all agencies involved there is huge room for improvement in the quality of the data we make available and how we make it accessible. One of *K-Lab's* tasks will be to cooperate on making quality data available through well documented API's and making sure as much data as possible have open licenses that allow unlimited re-use.

Piloting services. The work done in the last year of Heritage Here with the map service mentioned above demonstrated to all parties involved the importance of actually building services that make use of our own open data. *K-lab* will, as a part of its scope, function as a 'sandbox' for both coming up with new ideas for services, and – to the extent that budget and resources allow for it – try out new technologies and services. One such pilot service, is the work done by K-lab – in collaboration with the Estonian photographic heritage society – in setting up a crowdsourcing platform for improving metadata on historic photos (https://fotodugnad.ra.no/).

For 2018, *K-Lab* will start looking into building a service making use of linked open data from our organizations. All of our agencies are data owners that responsible for authority data in some form or another – ranging from geo names to cultural heritage data and person data. Some work has been done already to bring our technical departments closer in this field, but we plan to do 'something' on a practical level next year.

Building competence. In order to facilitate the exchange of knowledge between the collaboration partners *K-Lab* will arrange seminars, workshops and conferences as arenas for discussing common challenges, learning from each other and building networks. This is done primarily to strengthen the relationship between the agencies involved – but many activities will have a broader scope. One such example is the intention to arrange workshops – roughly every two months – on topics that are relevant to our agencies, but that are open to anyone interested. To give a rough overview of the range of topics, these workshops were arranged in 2017:

- A practical introduction to Cidoc-CRM (May)
- Workshop on Europeana 1914-1918 challenge – co-host: Wikimedia Norway (June)
- An introduction to KulturNAV – co-host: Vestfoldmuseene (September)
- Getting ready for #hack4no (October)

- Transkribus – Text recognition and transcription of handwritten text - co-host: The Munch museum (November)

Third phase – "Samarbeidsforum" 2018 and beyond

Towards the end of 2017 *K-lab* was very much a work in progress, and its future direction depended on many factors. However, a joint workshop was held in September 2017 to evaluate the work done so far – and to try and map out a way forwards for the collaboration. Employees from all levels in the organisations were present, with invited guests from other institutions from the cultural sector – like the National Library and Digisam from Sweden – to evaluate, discuss and suggest ideas.

No definite conclusions were drawn, but there was an overall agreement that the focus on the three areas described above is of great importance, and that the work done so far by the agencies together had been, for the most part, successful. Setting up arenas for discussing common problems, sharing success stories and interacting with colleagues across agency boundaries were regarded as key elements in the relative success of K-Lab so far. This work will continue into 2018 with focus on thematic groups on linked open data and photo archives, and a new series of workshops is being planned. The experimentation with technology will continue, and hopefully new ideas will be brought forward and realised over the course of the next year(s).

Towards the end of 2017 the basic premise of K-Lab changed somewhat due to 'external' conditions. Throughout 2017 the Ministry of Culture has been working on a strategy for open data, and the conclusions were made available towards of the end of the year (page is in Norwegian): https://www.regjeringen.no/no/dokumenter/kulturdepartementets-strategi-for-apne-data/id2576038/. One conclusion – and goal – was to "establish a forum for open cultural data – focusing on defining common needs and suggesting common solutions". And further: "the form of the cooperation will be decided, in cooperation, by The National Archives, The Arts Council and the National Library". A later version also specified that "The Directorate for Cultural Heritage is invited into the collaboration as a partner". At present, the most likely outcome of this open data strategy will be that K-Lab will most likely be replaced by the suggested cooperative forum – but details are yet to be decided.

Jānis Daugavietis, Rita Treija
Organisation(s): Institute of Literature, Folklore and Art - University of Latvia

Submitted by: Dr. Jānis Daugavietis (Institute of Literature, Folklore and Art - University of Latvia), ID: 1155 Topics: folklore and oral history, sociology, crowdsourcing, interdisciplinary collaboration
Keywords: online survey, methodology, questionnaires, folkloristics, sociology

CAWI for DH

ID 165

Abstract, final version, 2018.02.05.

The survey method of using questionnaires for acquiring different kinds of information from the population is a classic way to collect data. Mathematics, sociology and other sciences have developed a coherent theoretical methodology and have accumulated experience-based knowledge for using online survey tools, and the main research question in this paper concerns the differences between 'classical' (mostly quantitative) survey and those we are employing in Digital Humanities (DH): what are the most important aspects when it comes to CAWI (computer assisted web interview also called web, electronic, online, internet survey) for DH?

To answer this question, we will make a schematic comparison of most popular CAWI used in social sciences and those in DH, looking at previous experience of our work in fields and institutions of sociology, statistics and folkloristics, as well an analysis of recent DH literature.

Examples of surveys can be traced back to ancient civilizations, like censuses or standardised agricultural data recordings. The main instrument of this method is questioning (closed-ended or open-ended) which should be asked exactly the same way to all the representatives of surveyed population. During the last 20-25 years the internet survey method has been well developed and more and more frequently employed in social sciences and marketing research, among others. Usually CAWI is designed for acquiring quantitative data, but as in other most used survey modes (face-to-face, telephone or mail interviews) it can be used to collect qualitative data such as un- or semi-structured text/ speech, pictures, sounds etc.

In this paper, we focus on the case of Latvian folkloristics in the realm of DH**.** In recent years, the CAWI methodology has been used more and more within projects of the Institute of Literature, Folklore and Art of the University of Latvia (ILFA, UL). At the same time, the knowledge of this method in this field is somehow limited (because lack of previous experience and in many cases, education—the humanities curriculum usually does not include quantitative methods). This paper seeks to analyse specificity of CAWI designed for our needs of DH, which can differ significantly from those of the social sciences.

Questionnaires as an approach for collecting data of traditional culture date back to an early stage of the disciplinary history of Latvian folkloristics, namely, to the end of the 19th century and the beginning of the 20th century (published by Dāvis Ozoliņš, Eduard Wolter, Pēteris Šmits, Pēteris Birkerts). The Archives of Latvian Folklore was established in 1924. Its founder and the first Head, folklorist and schoolteacher Anna Bērzkalne, utilized questionnaires (*jautājumu lapas*) on various topics of Latvian folklore on a regular basis and distributed them to the Archives' contributors (schoolteachers, students, local historians and other volunteers). She both created original sets of questions herself and translated and adapted those by the Estonian and Finnish folklore scholars (instructions for collecting children's songs by Walter Anderson; questionnaires of folk beliefs by O. A. F. Mustonen alias Oskar Anders Ferdinand Lönnbohm and Viljo Johannes Mansikka). These localised equivalents were published in the press. Printed

questionnaires, such as “House and Household”, “Fishing and Fish”, “Relations between Relatives and Neighbors” and others, presented sets of questions of which were formulated in a suggestive way so that everyone who had some interest could easily engage in the work. The hand-written responses were sent to the Archives of Latvian Folklore from all regions of the country; the collection of folk beliefs in the late 1920s greatly supplemented the range of materials at the Archives.

However, the life of the survey as a method of collecting folklore in Latvia did not last long. Soon after World War II it was overcome by the dominance of professional collective fieldwork, and, at the end of the 20th century, by individual field research, including mainly face-to-face qualitative interviews with informants.

Only in 2017 did the Archives of Latvian Folklore, ILFA, UL, revitalize the approach of remote data collecting via the online questionnaires. With the project “Empowering knowledge society: interdisciplinary perspectives on public involvement in the production of digital cultural heritage” (funded by the European Regional Development Fund), a virtual inquiry module has been developed. The working group of virtual ethnography launched a series of online surveys aimed to study the calendric practices of individuals in the 21st century. Along with working out the iterative inquiry, data accumulation, and analysis tools, researchers have tried to find solutions to the technical and ethical challenges of the modern day.

The term meme was coined by Richard Dawkins, a British evolutionary biologist, in 1976 in his book *The Selfish Gene* as a unit of cultural transmission. Memes are a cultural electronic product that satirizes current popular events, and can be used to criticize those in power. The success of a meme is measured by its "virality" and the mutations that are reproduced from it like a germ or a part of the genetic trend of digital societies. I am interested in analyzing these new forms of the language of the internet in the context of the construction of the wall between the US and Mexico. I examine popular memes in Mexico and the US from both sides of the border. I believe these "political haikus" work as an escape valve for the tensions generated in the culture wars that consume American politics, particularly in the era of Trump. The border is an "open wound" that was opened after the War of 1847 and resulted in Mexico losing half of its territory. Currently, the wall functions as a political membrane barring the "expelled citizens" of south of the border from the economic benefits of the North. Memes help to expunge the gravity of a two-thousand-mile concrete wall in a region that shares cultural traits, languages, and natural environment, a region that cannot be domesticated with symbolic monuments to hatred. Memes are rhetorical devices that convey the absurdity of a situation, and can be a form of social participation as in a recent popular meme that shows a colorful Trojan horse-size piñata on the edge of the border, a meme that infantilizes the State-funded project of a fence. The meme's iconoclastography sets in motion a discussion of the real issues at hand—global economic disparities and the human planetary right to migrate.

Serious gaming to support stakeholder participation and analysis in Nordic climate adaptation research

Introduction

While climate change adaptation research has advanced significantly in recent years, we still lack a thorough discussion on maladaptation, i.e. the unintended negative outcomes as a result of implemented adaptation measures. In order to identify and assess examples of maladaptation for the agricultural sector, we developed a novel methodology, integrating visualization, participatory methods and serious gaming. This enables research and policy analysis of trade-offs between mitigation and adaptation options, as well as between alternative adaptation options with stakeholders in the agricultural sector. Stakeholders from the agricultural sector in Sweden and Finland have been engaged in the exploration of potential maladaptive outcomes of climate adaptation measures by means of a serious game on maladaptation in Nordic agriculture, and discussed their relevance and related trade-offs.

The Game

The Maladaptation Game is designed as a single player game. It is web-based and allows a moderator to collect the settings and results for each player involved in a session, store these for analysis, and display these results on a 'moderator screen'. The game is designed for agricultural stakeholders in the Nordic countries, and requires some prior understanding of the challenges that climate change can pose on Nordic agriculture, as well as the scope and function of adaptation measures to address these challenges.

The gameplay consists of four challenges, each involving multiple steps. At the start of the game, the player is equipped with a limited number of coins, which decrease for each measure that is selected. As such, the player has to consider the implications in terms of risk and potential negative effects of a selected measure as well as the costs for each of these measures. The player is challenged with four different climate related challenges – increased *precipitation*, drought, increased occurrence of *pests and weeds*, and a *prolonged growing season* - that are all relevant to Nordic agriculture. The player selects one challenge at a time. Each challenge has to be addressed, and once a challenge has been concluded, the player cannot return and revise the selection. When entering a challenge (e.g. *precipitation*) possible adaptation measures that can be taken to address this challenge in an agricultural context, are displayed as illustrated cards on the game interface. Each card can be turned to receive more information, i.e. a descriptive text and the related costs. The player can explore all cards before selecting one.

The selected adaptation measure then leads to a potential maladaptive outcome, which is again displayed as an illustrated card with an explanatory text on the backside. For each measure, there is a number of maladaptive outcomes which are selected at random for each individual game session. The player has to decide to reject or accept this potential negative outcome. If the maladaptive outcome is rejected, the player returns to the previous view, where all adaptation measures for the current challenge are displayed, and can select another measure, and make the decision whether to accept or reject the potential negative outcome that is presented for these. In order to complete a challenge, one adaptation measure with the related negative outcome has to be accepted. After completing a

challenge, the player returns to the entry page, where, in addition to the overview of all challenges, a small scoreboard summarizes the selection made, displays the updated amount of coins as well as a score of maladaptation-points. These points represent the negative maladaptation score for the selected measures and are a measure that the player does not know prior to making the decision.

The game continues until selections have been made for all four challenges. At the end of the game, the player has an updated scoreboard with three main elements: the summary of the selections made for each challenge, the remaining number of coins, and the total sum of the negative maladaptation score. The scoreboards of all players involved in a session appear now on the moderator screen. This setup allows the individual player to compare his or her pathways and results with other players. The key feature of the game is hence the stimulation of discussions and reflections concerning adaptation measures and their potential negative outcomes, both with regard to adding knowledge about adaptation measures and their impact as well as the threshold of when an outcome is considered maladaptive, i.e. what trade-offs are made within agricultural climate adaptation.

Analytical approaches to participant and game interaction

During autumn 2016, eight gaming workshops were held in Sweden and Finland. These workshops were designed as visualization supported focus groups, allowing for some general reflections, but also individual interaction with the web-based game. Stakeholders included farmers, agricultural extension officers, and representatives of branch organizations as well as agricultural authorities on the national and regional level. Focus group discussions were recorded and transcribed in order to analyze the empirical results with focus on participants´ interactions and meaning constructions of agricultural adaptation and potential maladaptive outcomes.

Given the multiple character of possible game session interactions, the analysis of the Maladaptation game differed between three types of interactions: 1) interactions between two or more co-located players, 2) interactions with narratives, images, and representations as expressed in the game or by the players, or 3) between culturally embedded traditions as expressed in and around the game.

Preliminary conclusions from the visualization supported gaming workshops

Preliminary conclusions from the visualization supported gaming workshops point towards several issues that relate both to content and functionality of the game. While, as a general conclusion, the stakeholders were able to quickly get acquainted with the game and interact without larger difficulties, some few individual participants were negative to the general idea of engaging with a game to discuss these issues. The level of interactivity that the game allows, where players can test and explore, before making a decision, enabled reflections and discussions also during the gameplay. Stakeholders frequently tested and returned to some of the possible choices before deciding on their final setting. While the game-player interaction allowed for a more individually oriented interaction, we found the combination of game - player interaction and player - player interaction to produce benefits in terms of communicative activities and richness in material. Hence, with the increase in digital and online games we anticipate an intensified discussion on the challenges of analyzing interactivity in digital gaming.

The combination of the three types of interaction, generated a large number of issues regarding the definition of maladaptive outcomes and their thresholds. The analysis found the game mediated research on climate change maladaptation to inform participant sense-making in relation to contextual aspects, such as temporal and spatial scales, as well as reflections regarding the relevance and applicability of the proposed adaptation measures and negative outcomes.

SuALT: Collaborative Research Infrastructure for Archaeological Finds and Public Engagement through Linked Open Data

Suzie Thomas[1], Anna Wessman[1], Jouni Tuominen[2,3], Mikko Koho[3], Esko Ikkala[3], Eero Hyvönen[2,3], Ville Rohiola[4], and Ulla Salmela[4]

[1] University of Helsinki, Dept. of Cultures, Finland
suzie.e.thomas@helsinki.fi, firstname.lastname@helsinki.fi
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://heldig.fi
[3] Semantic Computing Research Group (SeCo), Aalto University, Finland
http://seco.cs.aalto.fi, firstname.lastname@aalto.fi
[4] Finnish Heritage Agency, Finland
http://www.nba.fi, firstname.lastname@museovirasto.fi

The Finnish Archaeological Finds Recording Linked Database (Suomen arkeologisten löytöjen linkitetty tietokanta – SuALT) is a concept for a digital web service catering for discoveries of archaeological material made by the public; especially, but not exclusively, metal detectorists. SuALT, a consortium project funded by the Academy of Finland and commenced in September 2017, has key outputs at every stage of its development. Ultimately it provides a sustainable output in the form of Linked Data [3], continuing to facilitate new public engagements with cultural heritage [4], and research opportunities, long after the project has ended.

While prohibited in some countries, metal detecting is legal in Finland, provided certain rules are followed, such as prompt reporting of finds to the appropriate authorities and avoidance of legally-protected sites. Despite misgivings by some about the value of researching metal-detected finds, others have demonstrated the potential of researching such finds, for example uncovering previously unknown artefact typologies [1,6]. Engaging non-professionals with cultural heritage also contributes to the democratization of archaeology, and empowers citizens [2]. In Finland metal detecting has grown rapidly in recent years. In 2011 the Archaeological Collections registered 31 single or assemblages of stray finds. In 2014, over 2700 objects were registered, in 2015, near 3000. In 2016 over 2500 finds were registered. When the finds are reported correctly, their research value is significant. The Finnish Antiquities Act §16 obligates the finder of an object for which the owner is not known, and which can be expected to be at least 100 years old, to submit or report the object and associated information to The Finnish Heritage Agency (Museovirasto); the agency responsible for cultural heritage management in Finland. There is also a risk, as finders get older and even pass away, that their discoveries and collections will remain unrecorded and that all associated information is lost permanently.

In the current state of the art, while archaeologists increasingly use finds information and other data, utilization is still limited. Data can be hard to find, and available open data remains fragmented. SuALT will speed up the process of recording finds data. Because much of this data will be from outside of formal archaeological excavations,

it may shed light on sites and features not usually picked up through 'traditional' fieldwork approaches, such as previously unknown conflict sites. The interdisciplinary approach and inclusion of user research promotes collaboration among the infrastructure's producers, processors and consumers. By linking in with European projects, SuALT enables not only national and regional studies, but also contributes to international and transnational studies. This is significant for studies of different archaeological periods, for which the material culture usually transcends contemporary national boundaries. Ethical aspects are challenged due to the debates around engagement with metal detectorists and other artefact hunters by cultural heritage professionals and researchers, and we address head-on the wider questions around data sharing and knowledge ownership, and of working with human subjects. This includes the issues, as identified by colleagues working similar projects elsewhere, around the concerns of metal detectorists and other finders about sharing findspot information. Finally, the usability of datasets has to be addressed, considering for example controlled vocabulary to ease object type categorization, interoperability with other datasets, and the mechanics of verification and publication processes.

The project is unique in responding to the archaeological conditions in Finland, and in providing solutions to its users' needs within the context of Finnish society and cultural heritage legislation. While it focuses primarily on the metal detecting community, its results and the software tools developed are applicable more generally to other fields of citizen science in cultural heritage, and even beyond. For example, in many areas of collecting (e.g. coins, stamps, guns, or art), much cultural heritage knowledge as well as collections are accumulated and maintained by skillful amateurs and private collectors. Fostering collaboration, and integrating and linking these resources with those in national memory organizations would be beneficial to all parties involved, and points to future applications of the model developed by SuALT. Furthermore, there is scope to integrate SuALT into wider digital humanities networks such as DARIAH[5].

Framing SuALT's development as a consortium enables us to ask important questions even at development stages, with the benefit of expertise from diverse disciplines and research environments. The benefits of SuALT, aside from the huge potential for regional, national, and transnational research projects and international collaboration, are that it offers long term savings on costs, shares expertise and provides greater sustainability than already possible. We will explore the feasibility of publishing the finds data through international aggregation portals, such as Europeana[6] for cultural heritage content, as well as working closely with colleagues in countries that already have established national finds databases. The technical implementation also respects the enterprise architecture of Finnish public government. Existing Open Source solutions are further developed and integrated, for example the GIS platform Oskari[7] for geodata developed by the National Land Survey with the Linked Data based Finnish Ontology

[5] http://www.dariah.eu
[6] http://www.europeana.eu
[7] http://oskari.org

Service of Historical Places and Maps[8] [5]. SuALT's data is also disseminated through Finna[9], a leading service for searching cultural information in Finland.

SuALT consists of three subprojects:

1. Subproject "User Needs and Public Cultural Heritage Interactions" hosted by University of Helsinki;
2. Subproject "National Linked Open Data Service of Archaeological Finds in Finland" hosted by Aalto University,
3. Subproject "Ensuring Sustainability of SuALT" hosted by the Finnish Heritage Agency.

The primary aim of SuALT is to produce an open Linked Data service which is used by data producers (namely the metal detectorists and other finders of archaeological material), by data researchers (such as archaeologists, museum curators and the wider public), and by cultural heritage managers (FHA). More specifically, the aims are:

- To discover and analyse the needs of potential users of the resource, and to factor these findings into its development;
- To develop metadata models and related ontologies for the data that take into account the specific needs of this particular infrastructure, informed by existing models;
- To develop the Linked Data model in a way that makes it semantically interoperable with existing cultural heritage databases within Finland;
- To develop the Linked Data model in a way that makes it semantically interoperable with comparable 'finds databases' elsewhere in Europe, and
- To test the data resulting from SuALT through exploratory research of the datasets for archaeological research purposes for cultural heritage and collection management work.

The project corresponds closely with the strategic plans of the NBA and responds to the growth of metal detecting in Finland. Internationally, it corresponds with the development of comparable schemes in other European countries and regions, such as Flanders (MetaaldEtectie en Archeologie – MEDEA initiated in 2014), and Denmark and the Netherlands (Digitale Metaldetektorfund or DIgital MEtal detector finds – DIME, and Portable Antiquities in the Netherlands – PAN, both initiated in 2016). It takes inspiration from the Portable Antiquities Scheme (PAS) Finds Database[10] in England and Wales. These all aspire to an ultimate goal of a pan-European research infrastructure, and will work together to seek a larger international collaborative research grant in the future. A contribution of our work in relation to the other European projects is to employ the Linked Data paradigm, which facilitates better interoperability with related datasets, additional data enrichment based on well-defined semantics and reasoning, and therefore better means for analysing and using the finds data in research and applications.

[8] http://hipla.fi
[9] http://www.finna.fi
[10] https://finds.org.uk/database

References

1. Deckers, P.: “Productive” Sites in the Polders? “Griffin brooches” and Other Early Medieval Metalwork from the Belgian Coastal Plain. Medieval and Modern Matters 3, 21–43 (2012), https://doi.org/10.1484/J.MMM.5.102018
2. Dobat, A.S.: Between Rescue and Research: An Evaluation after 30 Years of Liberal Metal Detecting in Archaeological Research and Heritage Practice in Denmark. European Journal of Archaeology 16(4), 704–725 (2013)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/
4. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Morgan & Claypool, Palo Alto, CA (2012)
5. Ikkala, E., Tuominen, J., Hyvönen, E.: Contextualizing historical places in a gazetteer by using historical maps and linked data. In: Proc. of Digital Humanities 2016, short papers (2016)
6. Pedersen, U.: Viking Identities: Scandinavian Jewellery in England. Medieval History and Archaeology. Norwegian Archaeological Review 47(2), 225–227 (2014), https://doi.org/10.1080/00293652.2014.954609

The Dostoyevskian Trope
State Incongruence in Danish Textual Cultural Heritage

Kristoffer L Nielbo[1], Katrine F. Baunvig[2] and Jianbo Gao[3]

[1]Department of History, University of Southern Denmark, Odense, Denmark
[2]Study of Religions, Department of History, University of Southern Denmark, Odense, Denmark
[3]Institute of Complexity Science and Big Data Technology, Guangxi University, Nanning, China

Abstract

In the history of prolific authors, we are confronted with the figure of the suffering author. Setting aside metaphysical theories, the central claim seems to be that a state incongruent dynamic is an intricate part of the creativity process. Two propositions can be derived this claim, 1: the creative state is inversely proportional to the emotional state, and 2: the creative state is causally predicted by the emotional state. We call this creative-emotional dynamic 'The Dostoyevskian Trope'. In this paper we present a method for studying the Dostoyevskian trope in prolific authors. The method combines Shannon entropy as an indicator of lexical variability with fractal analysis in order to measure creative dynamics over multiple documents. We generate a sentiment time series from the same documents and test for asymmetric directed dependencies between the creative and sentiment time series. We illustrate the method by searching for the Dostoyevskian trope in Danish textual cultural heritage, specifically three highly prolific authors from the 19th century, namely, N.F.S. Grundtvig, H.C. Andersen, and S.A. Kierkegaard.

Keywords— Creativity, Fractal Analysis, Information Theory, Sentiment Analysis

Introduction

There is a popular belief among both laymen and researchers that suffering can promote creative production (1; 2). This belief has multiple names 'the suffering artist', 'the tortured artist' as well as 'the wisdom of suffering'. The validity of this belief is however debatable (3; 4). Setting aside various metaphysical theories about the clarity and humility inherent in pain, there is a central claim pertaining to the dynamic of the creative process, namely that, state incongruent dynamics are an intricate to the creativity process. We derive two testable propositions from this dynamic that we collectively call 'The Dostoyevskian Trope' with a reference to the paradigmatic life of Fodor Dostoyevsky: Firstly *the creative state is inversely proportional to the emotional state*, and secondly *the creative state is causally predicted by the emotional state*. If we can find empirical support both propositions, we argue that there is evidence favoring the Dostoyevskian trope. In order to

implement this, we propose a two step procedure: first, we combine information theory with fractal analysis in order to model creativity; second, we test for causal-like directed asymmetric relations between creativity and emotional states using average work sentiment as a proxy.

In this paper, creativity is modeled as *persistent trends in lexical variability*, estimated by the time varying Hurst exponent of Shannon's source entropy, for the collected writings of a prolific author. While entropy and information theory are widely used for studying variability in discrete time series such as characters and words (5; 6; 7), fractal analysis is less common in the humanities. Methods for fractal analysis are used in many areas of science to study self-similarity of complex dynamic systems (8). With self-similarity we simply mean that small parts of a system, in this case patterns of fluctuations at shorter time scales, are scaled copies of the larger parts of the system, that is, fluctuations at longer time scales. Reading, for instance, is a dynamic system that displays self-similarity across multiple time scales, because reading fluency and word comprehension are affected by the immediate word context (i.e. shorter time scales) as well as the larger text context (i.e., longer time scales) (9). Many culturally relevant complex systems display self-similarity in psychology (10), economy (11), sociology (12), and health (13), language (14) and music (15). In all these domains we find an important class of fractal objects called $1/f^{\alpha}$ noise, which is characterized by a power-law decaying power spectral density as well as power-law decaying rank-ordered eigenvalue spectrum. $1/f^{\alpha}$ noise has attracted considerably attention due to findings of so-called 'pink noise', where $\alpha = 1$, in a range of natural and man-made processes. In the present study we are targeting a subclass of $1/f^{\alpha}$ noise called a $1/f^{2H+1}$ process where H is the Hurst exponent that takes the values $0 < H < 1$. For $1/f^{2H+1}$ processes the following heuristic (16): For $0 < H < 0.5$, the time series is an anti-persistent process (i.e., increments are followed by decreases and decreases by increments), for $H = 0.5$, the time series only has short-range correlations also called short memory, and when $0.5 < H < 1$, the time series is a persistent process (i.e. increments are followed by increases and decreases by further decreases) characterized by long memory (Fig. 1). Returning the reading example, we can say that the reading is a persistent process because because it has been shown that for reading speed $0.5 < H < 1$.

With the massive data accumulation in almost every text domain, dictionary-based lexical matching has reemerged as a corpus independent approach to automated sentiment analysis. Sentiment content can be used to identify affective states and preferences of authors (17; 18; 19; 20), profile authors (21; 22), and construct narrative arcs (23; 24; 25). In accordance with this development, we use a Danish sentiment dictionary to estimate the average work sentiment in order to create a author-specific sentiment time series. This time series reflects fluctuations in expressive sentiment content relative to the individual author and can function as a limited proxy for the variability in his/her underlying emotional states. Multiple techniques can be used to compare asymmetric similarities between time varying creative and emotional states. In this paper we apply Granger causality (26), which tests for the existence and directionality of causal-like relations between temporally disjunctive time series. Granger causality, which originates in econometrics, is based on the assumption that causality is more than temporal disjunction, it involves directionality or predictability between time series. At its core Granger causality tests whether values of one time

series X contains information that is uniquely predictive of subsequent values in a different time series Y. The relation tested by Granger causality is often characterized as predictive causality and represented as $X\ Granger\ cause\ Y$ (27).

Methods

In this section we describe the method components of our approach focusing particularly on fractal analysis since it, in comparison to other analysis of lexical variability and sentiment, is less common in the humanities.

Entropy and Adaptive Fractal Analysis

In discrete cases with K distinct values, classical source entropy h of Shannon, or just entropy, is a measure of the variability. We measure h at the word level with a lexicon of K distinct types accordingly:

$$h = -\sum_{i=1}^{K} p_i \times \log_2(p_i) \tag{1}$$

With p being defined as:

$$p_i = Fr(w_i) / \sum_{i}^{K} Fr(w_i) \tag{2}$$

where Fr is the absolute frequency of word w. Word-level entropy measures the lexical variability of a tokenized string and will result in $0 \leqslant h \leqslant \log K$ with $h = \log K$ if the words in the string are uniformly distributed (i.e., a string where each type equally likely) and $h = 0$ if the string is perfectly predictable (i.e., a string with repetitions of only one type). For time series analysis in general, entropies are indicative of complexity such that larger values of h indicates greater complexity (28). Using entropy to model creativity in natural language is motivated by several studies (29; 30; 31; 7).

In fractal processes and time series analysis, detrended fluctuation analysis (DFA) (32) is a method for determining the statistical self-similarity (i.e. self-affinity) of a signal and can be used to analyze time series that appear to be long memory processes or $1/f$ noise. DFA is a widely used methods for estimating the Hurst parameter. It involves (1) constructing a random walk process

$$u(n) = \sum_{k=1}^{n} (x_k - \overline{x}), \quad n = 1, 2, \cdots, N, \tag{3}$$

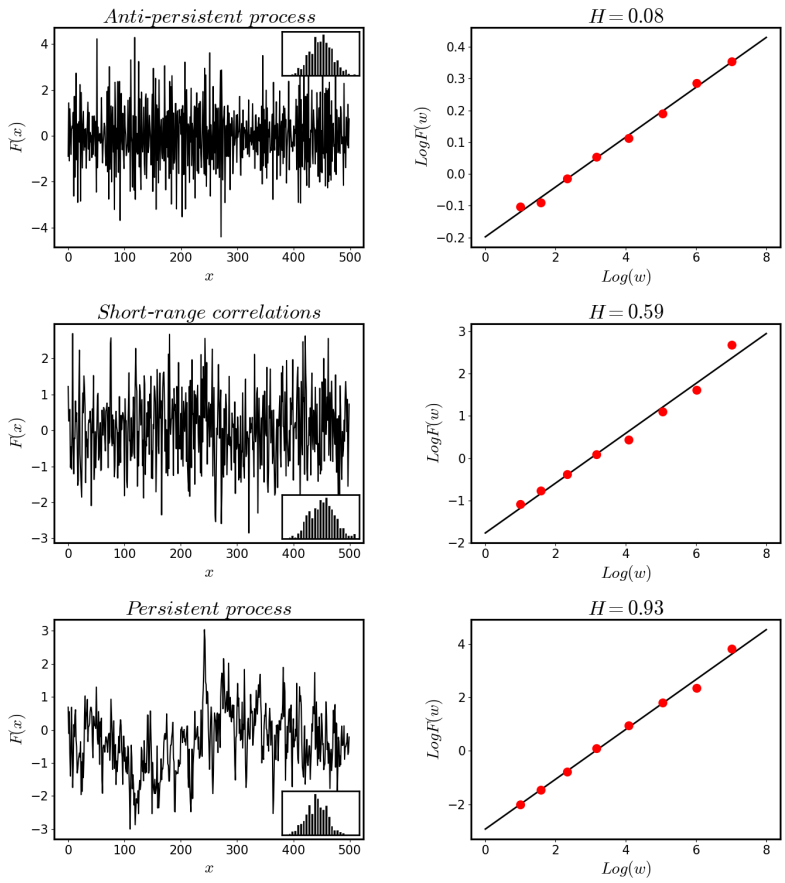


Fig. 1: Left: Time series that exhibit anti-persistent (top), short memory (middle), and persistent (bottom) behavior. Right: Estimation of the Hurst exponent for matching time series in the left column.

where $\overline{x}$ is the mean of the series $x(k)$, $k = 1, 2, \cdots, N$, (2) dividing the constructed random walk process into non-overlapping segments, (3) determining the best linear or polynomial fits in each segment as the local trends, (4) getting the variance of the differences between the random walk process and the local trends, and (5) averaging them over all the segments. Therefore, DFA may involve discontinuities at the boundaries of adjacent segments. Such discontinuities could be detrimental when the data contain trends (33), non-stationarity (34), or nonlinear oscillatory components such as signs of rhythmic activity (35; 36). To overcome this weakness adaptive fractal

analysis (AFA) has been proposed (16). AFA is an improvement of DFA. The main advantage of AFA over DFA is that AFA identifies a global smooth trend, which is obtained by optimally combining local linear or polynomial fitting, and thus no longer suffers from DFA's problem of discontinuities or even large, abrupt changes at the boundary of adjacent segments. As a result, AFA can automatically deal with arbitrary, strong nonlinear trends, which is not shared by any other methods, including DFA (16; 36).

AFA is based on a nonlinear adaptive multi-scale decomposition algorithm (16). The first step involves partitioning an arbitrary time series under study into overlapping segments of length $w = 2n + 1$, where neighboring segments overlap by $n + 1$ points. In each segment, the time series is fitted with the best polynomial of order M, obtained by using the standard least-squares regression; the fitted polynomials in overlapped regions are then combined to yield a single global smooth trend. Denoting the fitted polynomials for the $i - th$ and $(i + 1) - th$ segments by $y^i(l_1)$ and $y^{(i+1)}(l_2)$, respectively, where $l_1, l_2 = 1, \cdots, 2n + 1$, we define the fitting for the overlapped region as

$$y^{(c)}(l) = w_1 y^{(i)}(l + n) + w_2 y^{(i+1)}(l), \quad l = 1, 2, \cdots, n + 1, \tag{4}$$

where $w_1 = \left(1 - \frac{l-1}{n}\right)$ and $w_2 = \frac{l-1}{n}$ can be written as $(1 - d_j/n)$ for $j = 1, 2$, and where d_j denotes the distances between the point and the centers of $y^{(i)}$ and $y^{(i+1)}$, respectively. Note that the weights decrease linearly with the distance between the point and the center of the segment. Such a weighting is used to ensure symmetry and effectively eliminate any jumps or discontinuities around the boundaries of neighboring segments. As a result, the global trend is smooth at the non-boundary points, and has the right and left derivatives at the boundary (37). The global trend thus determined can be used to maximally suppress the effect of complex nonlinear trends on the scaling analysis. The parameters of each local fit is determined by maximizing the goodness of fit in each segment. The different polynomials in overlapped part of each segment are combined using Equation (3) so that the global fit will be the best (smoothest) fit of the overall time series. Note that, even if $M = 1$ is selected, i.e., the local fits are linear, the global trend signal will still be nonlinear. With the above procedure, AFA can be readily described. For an arbitrary window size w, we determine, for the random walk process $u(i)$, a global trend $v(i), i = 1, 2, \cdots, N$, where N is the length of the walk. The residual of the fit, $u(i) - v(i)$, characterizes fluctuations around the global trend, and its variance yields the Hurst parameter H according to the following scaling equation:

$$F(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{1/2} \sim w^H. \tag{5}$$

Thus, by computing the global fits, the residual, and the variance between original random walk process and the fitted trend for each window size w, we can plot $\log_2 F(w)$ as a function of $\log_2 w$. The presence of fractal scaling amounts to a linear relation in the plot, with the slope of the relation providing an estimate of H (Fig. 1).

Sentiment Analysis and Causal-like Dependencies

Dictionary-based (or lexicon-based) sentiment analysis is simple lexical matching utilizing a dictionary of words annotated for their sentiment value. Dictionaries are typically full form word lists and available dictionaries vary widely in terms of design, sentiment range, word classes, domains and languages [1]. Because analysis of social media and consumer behavior is an influential driver in the development of sentiment analysis, most dictionaries are developed for contemporary English. The present study uses the AFINN dictionary (38), which was developed contemporary Danish and English. In order to apply the dictionary to 19th century Danish, we normalized all text by 'modernizing' them using a combination of statistical and rule-based approaches to spelling correction. The average sentiment score for each work was then computed by averaging sentiment scores over slices of 1000 words (see Data).

Finally, a test for Granger causality was applied to sentiment time series and time-varying Hurst exponent for entropy. To test if variation in creativity (Y) at time t is predicted by emotional states (X) at earlier time steps $t-1 \ldots t-k$, that is if $X\ Granger\ cause\ Y$, we compared the nested ('creativity only') model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_k y_{t-k} + \epsilon \tag{6}$$

with the full ('creativity and emotional states') model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_k y_{t-k} + \alpha_1 x_{t-1} + \ldots + \alpha_k x_{t-k} + \epsilon \tag{7}$$

to see which one does the better job at explaining y_t based on th residuals. The zero-model for the hypothesis then is $H_0 : \alpha_i = 0$ for each i of the element $[1, k]$ with the alternative hypothesis being $H_1 : \alpha_i \neq 0$ for at least one i of the element $[1, k]$. We applied the test bi-directionally such that the Dostoyevskian Trope finds support iff we can confirm that '$X\ Granger\ cause\ Y$' and reject that '$Y\ Granger\ \ cause\ X$'.

Data

We illustrate the applicability of the method on a corpus ($N = 1329$ documents) by N.F.S. Grundtvig ($n_1 = 921$), H.C. Andersen ($n_2 = 194$), and S.A. Kierkegaard ($n_3 = 214$). Before analysis the corpus was reduced to alpha-numeric characters and each document was length normalized in windows of 1000 words.[2]. The entropy and sentiment scores for each document were then estimated as an average over windows. Time-varying estimation of H was computed maximally overlapping windows of entropy for 256 works.

[1]For an overview see: Reagan, A., Tivnan, B., Williams, J. R., Danforth, C. M., & Dodds, P. S. (2015). Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. ArXiv Preprint ArXiv:1512.00531.

[2]Normalizing to windows of 250 and 500 words results in qualitatively similar results.

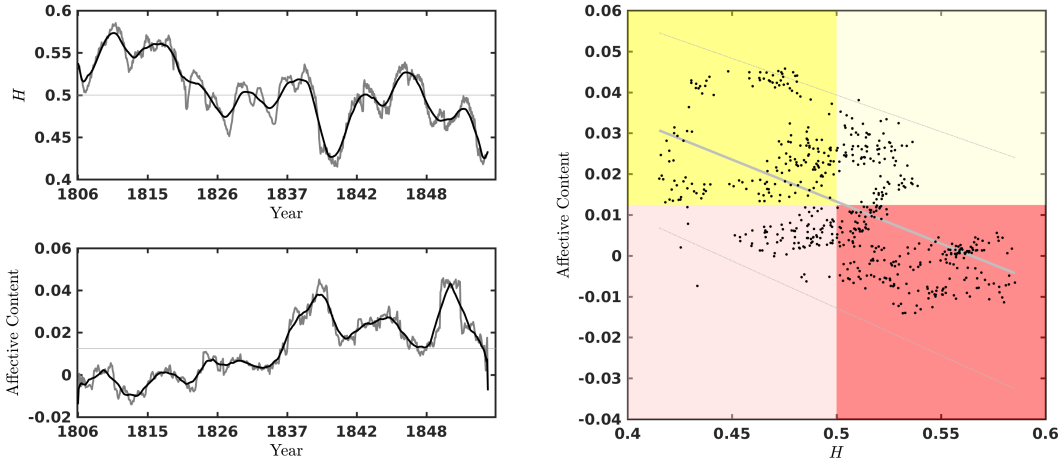


Fig. 2: Left: Time varying H(urst) exponent (upper) and sentiment (lower) for N.F.S. Grundtvig. Right: Inverse linear relation between H and sentiment (boundaries are 95Notice that only

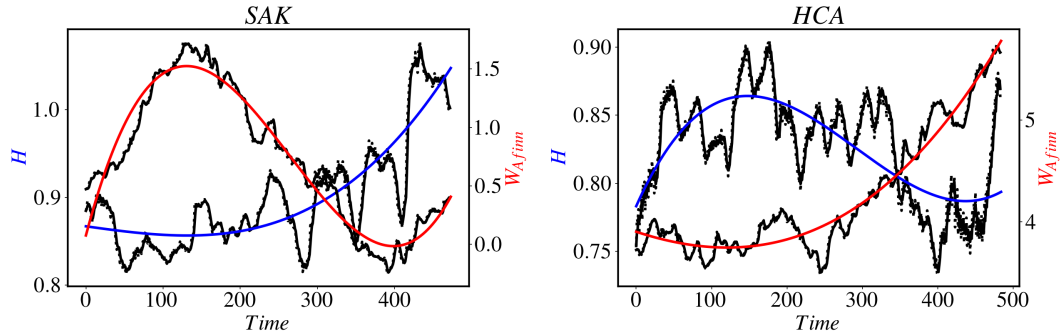


Fig. 3: Time varying H(urst) exponent (blue) and sentiment (red) for H.C. Andersen (HCA) (right) and S.A. Kierkegaard (SAK) (left). Red and blue lines are cubic fits to the raw signals. Both HCA and SAK show evidence of the first part of the Dostoyevskian trope, but only SAK displays a time delay in the hypothesized direction. HCA's Entropy exhibits persistent behavior ($0.5 < H_e \leq 1$) for his entire career

Results

We summarize the result as follows: First, all authors, N.F.S. Grundtvig (NFSG), H.C. Andersen (HCA) and S.A. Kierkegaard (SAK), show an inverse relation between creative and emotional states (Fig. 2 and 3). In accordance with previous studies on creativity and mood, this result is indicate an association between negative mood and artistic creativity (39; 40; 41). Second, only NFSG and

SAK show an asymmetric causal-like relation between in the hypothesized direction. We do in other words find support for the Dostoyevskian trope in Danish textual cultural heritage, but since HCA is propablt the most successful of the three, not for all authors have to be suffering[3] in order to be creatively successful.

For further research it is worth noticing that both NFSG and HCA display a general negative creativity trend, which might reflect aging. SAK died relatively young at the age of 42, which is roughly similar to the age at which NFSG and HCA's creativity decreases. Another interesting detail is that HCA's was creatively optimal in the sense that his creativity shows persistent trends throughout his career. NFSG's creatvity on the other hand becomes anti-persistent or rigid with age. SAK finally displays on-off intermittency for his later years switching between states of persistent behavior and "bursts" of chaotic behavior (42).

References

[1] K. R. Jamison, *Touched with Fire: Manic-Depressive Illness and the Artistic Temperament*, Free Press, New York, reissue edition edition, 1996.

[2] C. Zara, *Tortured Artists: From Picasso and Monroe to Warhol and Winehouse, the Twisted Secrets of the World's Most Creative Minds*, Adams Media, Avon, Massachusetts, 2012.

[3] L. A. King, L. M. Walker, and S. J. Broyles, Journal of research in personality **30**, 189 (1996).

[4] A. Furnham, D. J. Hughes, and E. Marshall, Thinking Skills and Creativity **10**, 91 (2013).

[5] C. E. Shannon, ACM SIGMOBILE Mobile Computing and Communications Review **5**, 3 (1948).

[6] P. Thoiron, Computers and the Humanities **20**, 197 (1986).

[7] Y. Zhang, Journal of Language Modelling **3**, 569 (2016).

[8] L. S. Liebovitch and L. A. Shehadeh, Tutrorials in contemporary nonlinear methods **24**, 178 (2003).

[9] B. A. O'Brien, S. Wallot, A. Haussmann, and H. Kloos, Scientific Studies of Reading **18**, 235 (2014).

[10] N. Chater and G. D. Brown, Cognition **69**, B17 (1999).

[11] T. Marchant, Social Choice and Welfare **31**, 693 (2008).

[12] J. Gao, P. Fang, and F. Liu, Physica A: Statistical Mechanics and its Applications **482**, 74 (2017).

[13] A. Eke, P. Herman, L. Kocsis, and L. R. Kozak, Physiological Measurement **23**, R1 (2002).

[3] With this method "suffering" can be translated to lower than life-time sentiment average.

[14] J. Gao, J. Hu, X. Mao, and M. Perc, Journal of The Royal Society Interface , rsif20110846 (2012).

[15] R. F. Voss and J. Clarke, Nature **258**, 317 (1975).

[16] J. Gao, J. Hu, and W.-w. Tung, PLoS ONE **6**, e24331 (2011).

[17] I. Fernández, D. Páez, and J. W. Pennebaker, International Journal of Clinical and Health Psychology **9** (2009).

[18] T. Nasukawa and J. Yi, Sentiment analysis: Capturing favorability using natural language processing, in *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, ACM, 2003.

[19] J. W. Pennebaker, Dynamics of Asymmetric Conflict **4**, 92 (2011).

[20] Y. R. Tausczik and J. W. Pennebaker, Journal of Language and Social Psychology **29**, 24 (2010).

[21] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, Text - Interdisciplinary Journal for the Study of Discourse **23**, 321 (2003).

[22] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, Effects of age and gender on blogging., in *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[23] P. S. Dodds et al., Proceedings of the National Academy of Sciences **112**, 2389 (2015).

[24] J. Gao, M. Jockers, J. Laudun, and T. Tangherlini, A multiscale theory for the dynamical evolution of sentiments in novels, 2017.

[25] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, arXiv preprint arXiv:1606.07772 (2016).

[26] C. W. Granger, Econometrica: Journal of the Econometric Society , 424 (1969).

[27] G. Sugihara et al., Science **338**, 496 (2012).

[28] C. Bandt and B. Pompe, Physical Review Letters **88** (2002).

[29] R. M. Bilder and K. S. Knudsen, Frontiers in Psychology **5** (2014).

[30] L. Gabora, A Possible Role for Entropy in Creative Cognition, page E001, MDPI, 2016.

[31] J. Rigau, M. Feixas, and M. Sbert, Conceptualizing Birkhoff's Aesthetic Measure Using Shannon Entropy and Kolmogorov Complexity., in *Computational Aesthetics*, pages 105–112, 2007.

[32] C.-K. Peng et al., Physical review e **49**, 1685 (1994).

[33] K. Hu, P. C. Ivanov, Z. Chen, P. Carpena, and H. Eugene Stanley, Physical Review E **64** (2001).

[34] J. W. Kantelhardt et al., Physica A: Statistical Mechanics and its Applications **316**, 87 (2002).

[35] Z. Chen et al., Phys. Rev. E **71**, 011104 (2005).

[36] J. Hu, J. Gao, and X. Wang, Journal of Statistical Mechanics: Theory and Experiment **2009**, P02066 (2009).

[37] M. A. Riley, S. Bonnette, N. Kuznetsov, S. Wallot, and J. Gao, Frontiers in Physiology **3** (2012).

[38] F. \. Nielsen, arXiv preprint arXiv:1103.2903 (2011).

[39] X. Hu and B. Yu, Exploring The Relationship Between Mood and Creativity in Rock Lyrics., in *ISMIR*, pages 789–794, 2011.

[40] K. Gasper, Journal of Experimental Social Psychology **39**, 248 (2003).

[41] M. Akinola and W. B. Mendes, Personality and Social Psychology Bulletin **34**, 1677 (2008).

[42] J. F. Heagy, N. Platt, and S. M. Hammel, Physical Review E **49**, 1140 (1994).

DHN 2018 Abstract: When Open becomes Closed: Findings of the Knowledge Complexity (KPLEX) Project.

Authors: ***Jennifer Edmond, Nicola Horsley, Elisabeth Huber, Georgina Nugent-Folan, Rihards Kalnins, Jörg Lehmann, Mike Priddy, Thomas Stodulka, and Andrejs Vasiljevs.***

Submitted and presented by:
Georgina Nugent-Folan (nugentfg@tcd.ie)

Relevant themes: Open Science, Cultural Heritage & The Future.

The future of cultural heritage seems to be all about "data." A Google search on the term "data" returns over 5.5 billion hits, but the fact that the term is so well embedded in contemporary discourse does not necessarily mean that there is a consensus as to what it is or should be. The lack of consensus regarding what data are on a small scale acquires greater significance and gravity when we consider that one of the major terminological forces driving ICT development today is that of "big data." So too do terms such as "data cleaning," "signal" and "noise." While the phrase may sound inclusive and integrative, "big data" approaches are highly selective, excluding any input that cannot be effectively structured, represented, or, indeed, digitised. The future of DH, of any approaches to understanding complex phenomena or sources such as are held in cultural heritage institutions, indeed the future of our increasingly datafied society, depends on how we address the significant epistemological fissures in our data discourse. This is not to say that digital data analysis approaches cannot also nurture epistemic multiplicity, but that there are observable biases to be found in some aspects of big data research. For example, how can researchers claim that "when we speak about data, we make no assumptions about veracity"[1] while one of the requisites of "big data" is "veracity"?[2] On the other hand, how can we expect humanities researchers to share their data on open platforms such as the European Open Science Cloud (EOSC) when we, as a community, resist the homogenisation implied and required by the very term "data," and share our ownership of it with both the institutions that preserve it and the individuals that created it? How can we strengthen European identities and transnational understanding through the use of ICT systems when these very systems incorporate and obscure historical biases between languages, regions, and power elites? In short, are we facing a future when the mirage of technical "openness" actually closes off our access to the perspectives, insight, and information we need as scholars and as citizens? How might this dystopic vision be avoided?

These are the questions and issues under investigation by the European Horizon 2020 funded Knowledge Complexity (KPLEX) project,[3] and we are doing so by applying strategies developed by humanities researchers to deal with complex, messy, cultural data; the very kind of data that resists datafication and poses the biggest challenges to knowledge creation in large data corpora environments. Arising out of the findings of the KPLEX project (the conclusion of which is fortuitously coterminous with DHN2018), this paper presents the

[1] Rosenberg, "Data before the Fact," in ibid., 37.
[2] See the cfp of BDIOT 2017 (http://www.bdiot.org/cfp.html) and IEEE 2017 (http://cci.drexel.edu/bigdata/bigdata2017/) respectively.
[3] The KPLEX project has been funded by the European Commission's Horizon 2020 Programme, Contract Number 732340.

synthesised findings of an integrated set of research questions and challenges addressed by a diverse team led by Trinity College Dublin (Ireland) and encompassing researchers in Freie Universität Berlin (Germany), DANS-KNAW (The Hague) and TILDE (Latvia). As this paper will make clear, we have adopted a comparative, multidisciplinary, and multi-sectoral approach to addressing the issue of bias in big data; focussing on the following four key challenges to the knowledge creation capacity of big data approaches:

1. Redefining what data is and the terms we use to speak of it (TCD);
2. The manner in which data that are not digitised or shared become "hidden" from aggregation systems (DANS-KNAW);
3. The fact that data is human created, and lacks the objectivity often ascribed to the term (FUB);
4. The subtle ways in which data that are complex almost always become simplified before they can be aggregated (TILDE).

This paper presents a synthesised version of these integrated research questions, combining qualitative and quantitative approaches to discuss the overall findings and recommendations of the project. What follows gives a flavour of the key issues addressed by the four project teams, and the related project findings that will be presented throughout this paper.

1. Redefining what data is and the terms we use to speak of it. Many definitions of data, even thoughtful scholarly ones, associate the term with a factual or objective stance, as if data were a naturally occurring phenomenon.[4] But data is not fact, nor is it objective, nor can it be honestly aligned with terms such as “signal” or “stimulus,” or the visceral but misleading “raw data.” To become data, phenomena must be captured in some form or agent, signal must be separated from noise and like must be organised against like. In short, transformations occur. These organisational processes are either human determined or human led, and therefore cannot be seen as wholly objective; irrespective of how effective a (human built) algorithm may be. The core concern of this keystone facet of the project was to expand extant understanding of the heterogeneity of definitions of data, and the implications of this state of understanding. By establishing a clear taxonomy of existing theories and definitions of data—identifying the key terms (and how they are used differently), key points of bifurcation, and key priorities under each conceptualisation of data—we provide a foundation which serves to underpin a more applied tautology of humanistic versus technical applications of the term. Our taxonomy of data definitions is backed up by the findings of a data mining exercise wherein we examined usage of the terms “data” and “big data” in the proceedings of major international big data journals from their inception to the present day. The overarching insights obtained by the taxonomy of data definitions and the data mining exercise are counterbalanced by the findings of thirteen in depth interviews with computer scientists, conducted with a view to obtaining a more detailed picture of the various understandings of the term “data” that underlie computer science research and development.

2. Dealing with “hidden” data. According to the 2013 ENUMERATE Core 2 survey, only 17% of the analogue collections of European heritage institutions had at that time been

[4] See for example Rowley, Jennifer. (2007) "The Wisdom Hierarchy: Representations of the DIKW Hierarchy." Journal of Information Science 33, no. 2: 163–80.

digitised.[5] This number actually represents a decrease over the findings of their 2012 survey (almost 20%). The survey also reached only a limited number of respondents: 1400 institutions over 29 countries, which surely captures the major national institutions but not local or specialised ones. Although the ENUMERATE Core 2 report does not break down these results by country, one has to imagine that there would be large gaps in the availability of data from some countries over others. Because so much of this data has not been digitised, it remains "hidden" from potential users. This may have always been the case, as there have always been inaccessible collections, but in a digital world the stakes and the perceptions are changing. The fact that so much other material is available on-line, and that an increasing proportion of the most well-used and well-financed cultural collections are available digitally as well, means that the reasonable assumption of the non-expert user of these collections is that what cannot be found does not exist (whereas in the analogue age, collections would be physically contextualised with their complements, leaving the more likely assumption to be that more information existed, but could not be accessed). The threat that our narratives of histories and national identities might thin out to become based on only the most visible sources, places and narratives is high. Through studying the often-neglected perspective of the archivist, this facet of the project elucidated the manner in which data that are not digitised or shared become "hidden" from aggregation systems and how cultural heritage practitioners are working to overcome the practical challenges that underlie ambitions of expanding access to public knowledge.

3. Knowledge organisation and the epistemics of emotions data. The nature of humanities data is such that even within the digital humanities, where research processes are better optimised toward the sharing of digital data, sharing of "raw data" remains the exception rather than the norm. The "instrumentation" of the humanities researcher consists of a dense web of primary, secondary and methodological or theoretical inputs, which the researcher traverses and recombines to create knowledge. This synthetic approach makes the nature of the data, even at its "raw" stage, quite hybrid, and already marked by the curatorial impulse that is preparing it to contribute to insight. This aspect may be more pronounced in the humanities than in other fields, but the subjective element is present in any human triggered process leading to the production or gathering of data. Emotions and affects serve as an effective cross-disciplinary topic, because few phenomena are as tricky in terms of datafication and measurement. This facet of the project elucidates that there is no shared vocabulary on either affect or emotion, nor data itself, with regards to differing "epistemic cultures."[6] By means of a series of in depth interviews and an online survey we investigated the researchers' view on affects' and emotions' resistance to datafication, and thus what escapes the access of positivistic approaches. Difficulties arise especially in the translation of scientific claims to objectivity[7] into research methodologies. Theoretical and methodological biases are inevitable when doing research on emotions and affects, and datafication therefore leads to a reduction and marginalization of the complexity of the research object "emotions." The insights gained will make visible many of the barriers to the inclusion of all aspects of science under current Open Science trajectories, and reveal further central elements of social and cultural knowledge that are unable to be

[5] http://www.enumerate.eu/en/statistics

[6] Karin Knorr-Cetina, Epistemic cultures: How the sciences make knowledge, Cambridge, Mass. u.a.: Harvard Univ. Press 1999.

[7] Lorraine Daston, Peter Galison, Objectivity, New York: Zone Books 2007.

accommodated under current conceptualisations of “data” and the systems designed to use them.

4. Cultural data and representations of system limitations. Cultural signals are ambiguous, polysemic, often conflicting, and contradictory. The process of “data-fication” robs culture of their polysemy, or at least reduces it to elements that have been, or can be, classified, divided, and filed into taxonomies and ontologies. One of the greatest challenges for so-called “big data” is the analysis and processing of multilingual content. This challenge is particularly acute for unstructured texts, which make up a large portion of the “big data” landscape. The language technology (LT) industry serves as an ideal test case for examining issues surrounding data, its availability, and the impact of data on technology, infrastructure, and employment. LT solutions are developed with language data as input material, therefore data issues—e.g., errors, noise, and inconsistencies in coverage—have a crucial impact on service quality. Input data issues become more acute when neural networks are used in development. AI-based solutions like Neural Machine Translation (NMT) are more sensitive to input data mistakes, often treating them as linguistic phenomena. Mistakes are exacerbated by data scarcity, particularly for smaller languages and overlooked domains. To understand the impact of data inconsistencies on LT, in this facet of the project we analyzed data availability for EU languages, including large-scale corpora, multilingual open data, and resources available for the CEF eTranslation platform, an analysis that has led to several hypotheses, subsequently validated by a survey of LT experts and EU language community representatives. In addition, we recount our efforts to analyze the effects of data inconsistencies on Neural MT. By exploring LT as a test case, we intended to show how data inequality will potentially become a major theme in “big data.” Furthermore, once it has been examined in the context of LT and “big data,” the overarching social and political consequences of data inequality will become apparent, helping to inform possibilities for policy decisions on the part of EU institutions.

Erik Edoff, Dept. of Culture and Media Studies, Umeå University, Sweden

A newspaper atlas: Named entity recognition and geographic horizons of 19th century Swedish newspapers

What was the outside world for 19th century newspaper readers? That is the overarching problem investigated in this paper. One way of facing this issue is to investigate what geographical places that was mentioned in the newspaper, and how frequently. For sure, newspapers were not the only medium that contributed to 19th century readers' notion of the outside world. Public meetings, novels, sermons, edicts, travelers, photography, and chapbooks are other forms of media that people encountered with a growing regularity during the century; however, newspapers often covered the sermons, printed lists of travelers and attracted readers with serial novels. This means, at least to some extent, that these are covered in the newspapers columns. And after all, the newspapers were easier to collect and archive than a public meeting, and thus makes it an accessible source for the historian.

Two newspapers, digitized by the National Library of Sweden, are analyzed: *Tidning för Wenersborgs stad och län* (TW) and *Aftonbladet* (AB). They are chosen based on their publishing places' different geographical and demographical conditions as well as the papers' size and circulation. TW was founded in 1848 in the town of Vänersborg, located on the western shore of lake Vänern, which was connected with the west coast port, Göteborg, by the Trollhätte channel, established in 1800. The newspaper was published in about 500 copies once a week (twice a week from 1858) and addressed a local and regional readership. AB was a daily paper founded in Stockholm in 1830 and was soon to become the leading liberal paper of the Swedish capital, with a great impact on national political discourse. For its time, it was widely circulated (between 5,000 and 10,000 copies) in both Stockholm and the country as a whole. Stockholm was an important seaport on the eastern coast. These geographic distinctions probably mean interesting differences in the papers' respective outlook. The steamboats revolutionized travelling during the first half of the century, but its glory days had passed around 1870, and was replaced by railways as the most prominent way of transporting people.

This paper is focusing on comparing the geographies of the two newspapers by analyzing the places mentioned in the periods 1848–1859 and 1890–1898. The main railroads of Sweden were constructed during the 1860s, and the selected years therefore cover newspaper geographies before and after railroads.

The main questions of paper addresses relate to media history and history of media infrastructure. During the second half of the 19th century several infrastructure technologies were introduced and developed (electric telegraph, postal system, newsletter corporations, railways, telephony, among others). The hypothesis is that these technologies had an impact on the newspapers' geographies. The media technologies enabled information to travel great distances in short timespans, which could have homogenizing effects on newspaper content, which is suggested by a lot of traditional research (Terdiman 1999). On the other hand, digital historical research has shown that the development of railroads changed the geography of Houston newspapers, increasing the importance of the near region rather than concentrating geographic information to national centers (Blevins 2014).

The goal of the study is in other words to investigate what these the infrastructural novelties introduced during the course of the 19th century as well as the different geographic and demographic conditions meant for the view of the outside world or the imagined geographies provided by newspapers. The primary goal with this paper is to investigate a historical-geographical problem relating to newspaper coverage and infrastructural change. The secondary aim is to tryout the use of Named Entity Recognition on Swedish historical newspaper data.

Named Entity Recognition (NER) is a software that is designed to locate and tag entities, such as persons, locations, and organizations. This paper uses SweNER to mine the data for locations mentioned in the text (Kokkinakis et al. 2014). Earlier research has emphasized the problems with bad OCR-scanning of historical newspapers. A picture of a newspaper page is read by an OCR-reading software and converted into a text file. The result contains a lot of misinterpretations and therefore considerable amount of noise (Jarlbrink & Snickars 2017). This is a big obstacle when working with digital tools on historical newspapers. Some earlier research has used and evaluated the performance of different NER-tools on digitized historical newspapers, also underlining the OCR-errors as the main problem with using NER on such data (Kettunen et al. 2017). SweNER has also been evaluated in tagging named entities in historical Swedish novels, where the OCR problems are negligible (Borin et al 2007). This paper, however, does not evaluate the software's result in a systematic way, even though some important biases have been identified by going through the tagging of some newspaper copies manually. Some important geographic entities are not tagged by SweNER at all (e.g. Paris, Wien [Vienna], Borås and Norge [Norway]). SweNER is able to pick up some OCR-reading mistakes, although many recurring ones (e.g. Lübeck read as Liibeck, Liibcck, Ltjbeck, Ltlbeck) are not tagged by SweNER. These problems can be handled, at least to some degree, by using "leftovers" from the data (wrongly spelled words) that was not matched in a comparison corpus. I have manually scanned the 50,000 most frequently mentioned words that was not matched in the comparative corpus, looking for wrongly spelled names of places. I ended up with a list of around 1,000 places and some 2,000 spelling variations (e.g. over 100 ways of spelling Stockholm). This manually constructed list could be used as a gazetteer, complementing the NER-result, giving a more accurate result of the 19th century newspaper geographies.

There is a big difference in size between the corpuses of two newspapers. As mentioned above, *AB* was published daily, which makes that corpus several times larger than the *TW* one. For the sample years I have analyzed (1850 and 1890), the corpus for *TW* is about 0,4 million words for 1850 and about 1,7 million for 1890. The preliminary result based on a sample from *TW* is shown in the table below. Regarding the precision of the NER location detection, some hints are given in the relation between manually tagged locations and locations tagged by NER, even though no formal precision and recall analysis has been performed yet.

Year	1850	1890
Corpus size (circa)	400,000 words	1,700,000 words
Locations tagged by NER	1,963	19,914
Locations tagged manually	1,359	8,129
Locations (total)	3,322	28,043
Unique locations	305	1,206
Locations/100,000 words	830	1,171

REFERENCES

Blevins, C. (2014), ”Space, nation, and the triumph of region: A view on the world from Houston”, *Journal of American History*, Vol. 101, no 1, pp. 122–147.

Borin, L., Kokkinakis, D., and Olsson, L-G. (2007), “Naming the past: Named entity and animacy recognition in 19th century Swedish literature”, *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pp. 1–8, available at: http://spraakdata.gu.se/svelb/pblctns/W07-0901.pdf (accessed October 31 2017).

Jarlbrink, J. and Snickars, P. (2017), “Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive”, *Journal of Documentation*, Vol. 73, no 6, pp. 1228–1243.

Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., and Löfberg, L. (2017), ”Old content and modern tools: Searching named entities in a Finnish OCRed historical newspaper collection 1771–1910”, *Digital Humanities Quarterly*, (preview) Vol. 11, no 3.

Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., and Borin, L., (2014), ”HFST-SweNER – A new NER resource for Swedish”, *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik 26–31 May 2014., pp. 2537-2543

Terdiman, R. (1999) “Afterword: Reading the news”, *Making the news: Modernity & the mass press in nineteenth-century France*, Dean de la Motte & Jeannene M. Przyblyski (eds.), Amherst: University of Massachusetts Press.

Spheres of "public" in eighteenth-century Britain

Mark J. Hill (presenter), Antti Kanner, Jani Marjanen, Ville Vaara, Leo Lahti, Eetu Mäkelä, Mikko Tolonen

Revised Abstract for Nordic Digital Humanities Conference in Helsinki, March 2018

The eighteenth-century saw a transformation in the practices of public discourse. With the emergence of clubs, associations, and, in particular, coffee houses, civic exchange intensified from the late seventeenth century. At the same time print media was transformed: book printing proliferated; new genres emerged (especially novels and small histories); works printed in smaller formats made reading more convenient (including in public); and periodicals - generally printed onto single folio half-sheets - emerged as a separate category of printed work which was written specifically for public consumption, and with the intention of influencing public discourse (such periodicals were intended to be both ephemeral and shared, often read, and then discussed, publically each day). This paper studies how these changes may be recognized in language by quantitatively studying the word "public" and its semantic context in the Eighteenth-Century Collections Online (ECCO).

While there are many descriptions of the transformation of public discourse (both contemporary and historical), there has been limited research into the language revolving (and evolving) around "public" in the eighteenth-century. Jürgen Habermas (2003: 2-3) famously argues that the emergence of words such as "*Öffentlichkeit*" in German and "publicity" in English are indicative of a change in the public sphere more generally. The conceptual history of "*Öffentlichkeit*" has been further studied in depth by Lucian Hölscher (1978), but a systematic study of the semantic context of "public" in British eighteenth-century material is missing. Studies that have covered this topic, such as Gunn (1989), base their findings on a very limited set of source material. In contrast, this study, by using a large-scale digitized corpus, aims to supplement earlier studies that focus on individual speech acts or particular collections of sources, and provide a more comprehensive account of how the language of "public" changed in the eighteenth century.

The historical subject matter means that the study is based on the ECCO corpus. While ECCO is in many ways an invaluable resource, a key goal of this study is to be methodologically sound from the perspective of corpus-linguistics and intellectual history, while developing insights which are relevant more generally to sociologists and historians. In this regard, ECCO does come with its own particular problems: both in terms of content and size.

With regard to content: OCR mistakes remain problematic; its heterogeneity in genres can skew investigations; and the unpredictable nature of duplicate texts introduced by numerous reprints of certain volumes must be taken into account. However, many of these problems can be mitigated in different ways. For example, in specific cases we compare findings with the, much smaller, ECCO TCP (an OCR corrected subset of ECCO). We have further used the English Short Title Catalogue (ESTC) to connect textual findings with relevant metadata information contained in the catalogue. By merging ESTC metadata with ECCO, one can

more easily use existing historical knowledge (for example, issues around reprints and multiple editions) to engage with the corpus.

With regard to size: the corpus itself is too big to run with automatic parsers (in terms of computing time and resources). We have therefore extracted a separate, and smaller, corpus (with the help of ESTC metadata) to do more complex and demanding analyses. The subcorpus is roughly 0.2% the size of the original corpus (this number remains relatively stable for each decade of the century), yet is still made of 19,945,971 tokens. Additionally, results of analyses on the sub-corpora were replicated (in a much simpler and cruder form) on the whole dataset, offering results which corroborate initial observations.

The size constraints provide their own advantages, however. The smaller subsections were chosen to represent pamphlets and other similar short documents by extracting all documents with less than 10406 characters in them. Compared to other specific genres or text types, this proved to be a successful method when attempting to define a meaningful subcorpus. Advantages include: limiting effects of reprints; including a relatively large number of individual writers in the analysis; subjects covered by pamphlets tend to be historically topical, and thus cleaner evidence for claims in terms of diachronic meaning; and as shorter texts, inspecting single occurrences in their original context is much more efficient as things such as theme, context, and writer's intentions reveal themselves comparatively quickly compared to larger works. Thus, issues around distant and close reading are more easily overcome. In addition, we are able to compare semantic change between the larger corpus and the more rapidly shifting topical and political debates found in pamphlets, which offers its own historical insights.

In terms of specific linguistic approaches, analysis started with examinations of contextual distributions of "public" by year. Then, by changing the parameters of this analysis (for example, by defining the context as a set of syntactic dependencies engaged by public, or as collocation structures of a wider lexical environment) different aspects of the use of "public" can be brought to the foreground.

As syntactic constraints govern possibilities of combinations of words in shorter ranges of context, the narrower context windows contain a lot of syntactic information in addition to collocational information. Because of this syntactic restrictedness of close range combinations, the semantic relatedness of words with similar short range context distributions is one of degree of mutual interchangeability and, as such, of metaphorical relatedness (Heylen, Peirsman, Geeraerts, Speelman 2008). Wider context windows, such as paragraphs, are free from syntactic constraints, and so semantic relatedness between two words with similar wide range context distributions carries information from frequent contiguity in context and can be described as more metonymical than metaphorical by nature, as is visible from applications based on term-document-matrices, such as topic modelling or Latent Semantic Analysis (cf. Blei, Ng and Jordan (2003) and Dumais (2005))

The syntactic dependencies were counted by analysing the pamphlet subcorpus using Stanford Lexical Parser (Cheng and Manning 2014). Results show changes in the tendency to use "public" as an adjective attribute and in compound positions. Since in English the overwhelmingly most frequent position for both adjective attributes and compounding attributes is preceding head words, this analysis could be adequately replicated using bigrams in the whole dataset. Lexical environments have been analysed by clustering second order collocations (cf. Bertels and Speelman (2014)) and replicated by using a random sampling from the whole dataset to produce the second order vectors.

The study of all bigrams relating to "public" (such as "public opinion", "public finances", "public religion") in ECCO provides for a broader analysis of the use of "public" in eighteenth-century discourse that not only focuses on particular compounds, but provides a better idea of which domains "public" was used in. It points towards a declining trend in relative frequency of religious bigrams during the course of the eighteenth century and rise in the relative frequency of secular bigrams - both political and economic. This allows us to present three arguments: First, it is argued that this is indicative of an overall shift in the language around "public" as the concept's focus changed and it began to be used in new domains. This expansion of discourses or domains in which "public" was used is confirmed in the analyses of a wider lexical environment. Second, we also notice that some collocates to public, such as "public opinion" and "public good", gained a stronger rhetorical appeal. They became tropes in their own right and gained a future orientation in political discourse in the latter half of the eighteenth century (Koselleck 1972). Third, by combining the results of the distributional semantics of "public" in ECCO with information extracted from ESTC, one can recognize how different groups used the language relating to "public" in different ways. For example, authors writing on religious topics tended to use "public" differently from authors associated with the enlightenment in Scotland or France.

There are two important upshots to this study: the methodological and the historical. With regard to the former, the paper works as a convincing case study which could be used as an example, or workflow, for studying other words that are pivotal to large structural change. With regard to the latter, the work is of particular historical relevance to recent discussions in eighteenth century intellectual history. In particular, the study contributes to the critical discussion of Habermas that has been taking place in the English-speaking world since the translation of his *Structural Transformation of the Public Sphere* in 1989, while also informing more traditional historical analyses which have not been able to draw tools from the digital humanities (Hill 2017).

References

Bertels, Ann and Dirk Speelman (2014). "Clustering for semantic purposes. Exploration of semantic similarity in a technical corpus." *Terminology* 20:2, pp. 279–303. John Benjamins Publishing Company.

Blei, David, Andrew Y. Ng and Michael I. Jordan (2003). "Latent Dirichlecht Allocation." *Journal of Machine Learning Research* 3 (4–5). Pp. 993–1022.

Chen, Danqi and Christopher D Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks." *Proceedings of EMNLP 2014*.

Dumais, Susan T. (2005). Latent Semantic Analysis. *Annual Review of Information Science and Technology*. 38: 188–230.

Gunn, J.A.W. (1989). "Public opinion.' *Political Innovation and Conceptual Change* (Edited by Terence Ball, James Farr & Rusell L. Hanson). Cambridge: Cambridge University Press.

Habermas, Jürgen (2003 [1962]). *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Cambridge: Polity.

Heylen, Christopher, Yves Peirsman, Dirk Geeraerts and Dirk Speelman (2008). "Modelling Word Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms." *Proceedings of LREC 2008*.

Hill, Mark J. (2017), "Invisible interpretations: reflections on the digital humanities and intellectual history." *Global Intellectual History* 1.2, pp. 130-150.

Hölscher, Lucian (1978), "'Öffentlichkeit.'" Otto Brunner et al. (Hrsg.) *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Band 4, Stuttgart, Klett-Cotta, pp. 413–467.

Koselleck, Reinhart (1972), "'Einleitung.'" Otto Brunner, Werner Conze & Reinhart Koselleck (hrsg.), *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Band I, Stuttgart, Klett-Cotta, pp. XIII–XXVII.

Jan Torpus, jan.torpus@fhnw.ch
University of Applied Sciences and Arts Northwestern Switzerland (FHNW), Institute of Experimental Design and Media Cultures (IXDM)

Extending museum exhibits by embedded media content for an embodied interaction experience

Investigation topic

Nowadays, museums not only collect, categorize, preserve and present; a museum must also educate and entertain, all the while following market principles to attract visitors. To satisfy this mission, they started to introduce interactive technologies in the 1990s, such as multimedia terminals and audio guides, which have since become standard for delivering contextual information. More recently there has been a shift towards the creation of personalized sensorial experiences by applying user tracking and adaptive user modeling based on location-sensitive and context-aware sensor systems with mobile information retrieval devices. However, the technological gadgets and complex graphical user interfaces (GUIs) generate a separate information layer and detach visitors from the physical exhibits. The attention is drawn to the screen and the interactive technology becomes a competing element with the environment and the exhibited collection [Stille 2003, Goulding 2000, Wakkary 2007]. Furthermore, the vast majority of visitors comes in groups and the social setting gets interrupted by the digital information extension [Petrelli 2016]. Exhibitions generate encounters of the visitor's lifeworld with the exhibits' objectworld [Wood 2016]. Objects contain information about material and physical characteristics, functionalities, actions and events, cultural and historical context and associated people during their entire lifespan. These aspects can be extended by the allocated media extensions that are subtly staged within the exhibition space. First studies about museum visitor behavior were carried out at the end of the 19th and during the 20th Century [Robinson 1928, Melton 1972]. More recently, a significant body of ethnographic research about visitor experience of single persons and groups has contributed studies about technologically extended and interactive installations. Publications about visitor motivation, circulation and orientation, engagement, learning processes, as well as cognitive and affective relationship to the exhibits are of interest for our research approach [Bitgood 2006, Vom Lehn 2007, Dudley 2010, Falk 2011]. Most relevant are studies of the Human Computer Interaction (HCI) researcher community in the fields of Ubiquitous Computing (ubiComp), Tangible User Interfaces and Augmented Reality (AR), investigating hybrid exhibition spaces and the bridging of the material and physical with the technologically mediated and virtual [Hornecker 2006, Wakkary 2007, Benford 2009, Petrelli 2016].

Approach

At the Institute of Experimental Design and Media Cultures (IXDM) we have conducted several design research projects applying AR for cultural applications but got increasingly frustrated with disturbing GUIs and physical interfaces such as mobile phones and Head Mounted Displays. We therefore started to experiment with Ubiquitous Computing, the Internet of Things and physical computing technologies that became increasingly accessible for the design community during the last twelve years because of shrinking size and price of sensors, actuators and controllers. In the presented research project, we therefore examine the extension of museum exhibits by physically embedded media technologies for an embodied interaction experience. We intend to overcome problems of distraction, isolation and stifled learning processes with artificial GUIs by interweaving mediated information directly into the context of the exhibits and by triggering events according to visitor behavior.

Our research approach was interdisciplinary and praxis-based including the observation of concept, content and design development and technological implementation processes before the final

evaluations. The team was composed of two research partners, three commercial/engineering partners and three museums, closely working together on three tracks: technology, design and museology. The engineering partners developed and implemented a scalable distributed hardware node system and a Linux-based content management system. It is able to detect user behavior and accordingly process and display contextual information. The content design team worked on three case studies following a scenario-driven prototyping approach. They first elaborated criteria catalogues, suitable content and scenarios to define the requirement profiles for the distributed technological environment. Subsequently, they carried out usability studies in the *Critical Media Lab* of the IXDM and finally set up and evaluated three case studies with test persons. The three museums involved, the *Swiss Open-Air Museum Ballenberg,* the *Roman City of Augusta Raurica and* the *Museum der Kulturen Basel*, all have in common that they exhibit objects or rooms that function as staged knowledge containers and can therefore be extended by means of ubiComp technologies.



Figure 1. *Roman City of Augusta Raurica*, case study: *"The Roman trade center Schmidmatt"*.
Figure 2. *Open-Air Museum Ballenberg*, case study: *"The farmhouse Uesslingen".*
Figure 3. *Museum der Kulturen Basel*, case study: "*Meditation box*".

The three case studies were thematically distinct and offered specific exhibition situations:

- Case study 1: *Roman City of Augusta Raurica: "The Roman trade center Schmidmatt*". The primary imparting concept was "oral history", and documentary film served as a related model: An archaeologist present during the excavations acted as a virtual guide, giving visitors information about the excavation and research methods, findings, hypotheses and reconstructions.



Figure 4. Prototypical catwalk system for test visitors.
Figure 5. Test visitor with video projection and illuminated replica.
Figure 6. Projection mapping onto a hypocaust floor and wall allows "x-ray view" to understand the construction.

- Case study 2: *Open-Air Museum Ballenberg:* "Farmhouse from Uesslingen". The main design investigation was "narratives" about the former inhabitants and the main theme "alcohol": Its use for cooking, medical application, religious rituals and abuse.



Figure 7. Sensors and nodes are hidden in the furniture.
Figure 8. Kitchen with video projection onto book and scenic sounds.
Figure 9. Bedroom with responsive video projected stains and illuminated medical utensils.

- Case study 3: *Museum der Kulturen Basel:* "Meditation box". The main design investigation was "visitor participation" with biofeedback technologies.



Figure 10. Usability study setup at *IXDM's Critical Media Lab*.
Figure 11. Visitor evaluation setup: Sofa (containing main technology items), touch-sensitive handle (3D printed lotus pedestal) and biofeedback chest belt visitors can wear.
Figure 12. Mandala behind semi-transparent textile with projected video animation explaining its functions.

Technological development

This project entailed the development of a prototype for a commercial hardware and software toolkit for exhibition designers and museums. Our technology partners elaborated a distributed system that can be composed and scaled according to the specific requirements of an exhibition. The system consists of two main parts:

- A centralized database with an online content management system (CMS) to setup and control the main software, node scripts, media content and hardware configuration. After the technical installation it also allows the museums to edit, update, monitor and maintain their exhibitions.
- Different types of hardware nodes that can be extended by specific types of sensors and actuators. Each node, sensor and actuator has its own separate ID; they are all networked together and are therefore individually accessible via the CMS. A node can run on a Raspberry Pi, for example, an FPGA based on Cyclone V or any desktop computer and can thus be adapted to the required performance.

The modular architecture allows for technological adaption or extension according to specific needs. First modules were developed for the project and then implemented according to the case study scenarios.

Evaluation methods

Through a participatory design process, we developed a scenario for each case study, suitable for walkthrough with several test persons. Comparable and complementary case study scenarios allowed us to identify risks and opportunities for exhibition design and knowledge transfer and define the tasks and challenges for technical implementation. For the visitor evaluation, we selected end-users, experts and in-house museum personnel. The test persons were of various genders and ages (including families with children), had varying levels of technical understanding and little or no knowledge about the project. For each case study we asked about 12 persons or groups of persons to

explore the setting as long as they wanted (normally 10–15 minutes). They agreed to be observed and video recorded during the walkthrough and to participate in a semi-structured interview afterwards. We also asked the supervisory staff about their observations and mingled with regular visitors to gain insight into their primary reactions, comments and general behavior. The evaluation was followed by a heuristic qualitative content analysis of the recorded audio and video files and the notes we took during the interviews. Shortly after each evaluation we presented and discussed the results in team workshops.

Findings and Conclusions

The field work lead to many detailed insights about interweaving interactive mediated information directly into the context of physical exhibits. The findings are relevant for museums, design researchers and practitioners, the HCI community and technology developers. We organized the results along five main investigation topics:

1. Discovery-based information retrieval

Unexpected ambient events generate surprise and strong experiences but also contain the risk of information loss if visitors do not trigger or understand the media aids. The concept of unfolding the big picture by gathering distributed, hidden information fragments requires visitor attentiveness. Teasing, timing and the choice of location are therefore crucial to generate flowing trajectories.

2. Embodied interaction

The ambient events are surprising but visitors are not always aware of their interactions. The unconscious mode of interaction lacks of an obvious interaction feedback. But the fact that visitors do not have to interact with technical devices or learn how to operate graphical user interfaces means that no user groups are excluded from the experience and information retrieval.

3. Non-linear contextual information accumulation

When deploying this project's approach as a central exhibition concept, information needs to be structured hierarchically. Text boards or info screens are still a good solution for introducing visitors to the ways they can navigate the exhibition. The better the basic topics and situations are initially introduced, the more freedom emerges for selective and memorable knowledge staged in close context to the exhibits.

4. Contextually extended physical exhibits

A crucial investigation topic was the correlation between the exhibit and the media extension. We therefore declined concepts that would overshadow the exhibition and would use it merely as a stage for storytelling with well-established characters or as an extensive media show. The museums requested that media content fades in only shortly when someone approaches a hotspot and that there are no technical interfaces or screens for projections that challenge the authenticity of the exhibits. We also discussed to what extend the physical exhibit should be staged to bridge the gap to the media extension.

5. Invisibly embedded technology

The problem of integrating sensors, actuators and controllers into cultural heritage collections was a further investigation topic. We used no visible displays to leave the exhibition space as pure as possible and investigated the applicability of different types of media technologies.

Final conclusion

Our museum partners agreed that our approach should not be implemented as a central concept and dense setting for an exhibition. If ubiComp is applied for discovery-based embodied interaction displaying contextual information without hierarchical structures, the approach should only be applied as a discreet additional information layer or just as a tool to be used when it makes sense to

explain something contextually or involve visitors emotionally. However, the developed sensor-actor-network and the database with a CMS for setup and maintenance also allows ubiComp implementations that are suitable for an overall concept of interactive information presentation: Areas for embodied interaction could be indicated, sensor activity could trigger distinguishable feedback, audiovisual displays could be integrated as aesthetic electronic devices to present hierarchically structured information and to help visitors to orientate and to retrieve all offered information. But in our opinion under these conditions embodied interaction would make no sense and devices with GUIs or even simple buttons would be more appropriate for visitor interaction.

References

Steve Benford et al. 2009. From Interaction to Trajectories: Designing Coherent Journeys Through User Experiences. Proc. CHI '09, ACM Press. 709–718.

Stephen Bitgood. 2006. An Analysis of Visitor Circulation: Movement Patterns and the General Value Principle. Curator the museum journal, Volume 49, Issue 4,463–475.

John Falk. 2011. Contextualizing Falk's Identity-Related Visitor Motivational Model. Visitors Studies. 14, 2, 141-157.

Sandra Dudley. 2010. Museum materialities: Objects, sense and feeling. In Dudley, S. (ed.) Museum Materialities: Objects, Engagements, Interpretations. Routledge, UK, 1-18.

Christina Goulding. 2000. The museum environment and the visitor experience. European Journal of marketing 34, no. 3/4, pp. 261-278.

Eva Hornecker and Jacob Buur. 2006. Getting a Grip on Tangible Interaction: A Framework on Physical Space and Social Interaction. CHI, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 437-446.

Dirk vom Lehn, Jon Hindmarsh, Paul Luff, Christian Heath. 2007. Engaging Constable: Revealing art with new technology. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), 1485-1494.

Arthur W. Melton. 1972. Visitor behavior in museums: Some early research in environmental design. In Human Factors. 14(5): 393-403.

Edward S. Robinson. 1928. The behavior of the museum visitor. Publications of the American Association of Museums, New Series, Nr. 5. Washington D.C.

Daniela Petrelli, Nick Dulake, Mark T. Marshall, Anna Pisetti, Elena Not. 2016. Voices from the War: Design as a Means of Understanding the Experience of Visiting Heritage. Proceedings Human-Computer Interaction, San Jose, CA, USA.

Alexander Stille. 2003. The future of the past. Macmillan. Pan Books Limited.

Ron Wakkary and Marek Hatala. 2007. Situated play in a tangible interface and adaptive audio museum guide. Published online: 4 November 2006. Springer-Verlag London Limited.

Elizabeth Wood and Kiersten F Latham. 2016. The Objects of Experience: Transforming Visitor-Object Encounters in Museums. Routledge, New York, USA.

PhD Sari Östman
MA Elina Vaahensalo
PhD Riikka Turtiainen
Digital Culture
University of Turku

DHN2018 conference / Abstract for short paper

Where are you going, research ethics in Digital Humanities?

1 Background

In this paper we will examine *the current state and possibilities for future development of research ethics among Digital Humanities*. This is an ongoing project, the results of which will be shown by the end of 2018. We have and will be analyzing the following:

a) ethics-focused inquiries with researchers in a multidisciplinary consortium project (CM)
b) Digital Humanities -oriented journals and
c) the objectives of the DigiHum Programme at the Academy of Finland, ethical guidelines of AoIR[1] and academical ethical boards and committees, in particular the one at the University of Turku.

Östman and Vaahensalo work in the consortium project *Citizen Mindscapes* (CM), which is part of the Academy of Finland's Digital Humanities Programme. University Lecturer Turtiainen is using a percentage of her work time for the project.

In the Digital Humanities program memorandum, *ethical examination of the research field* is mentioned as one of the main objectives of the program (p. 2). The CM project has a work package for researching research ethics, which Östman is leading. We aim at examining the current understanding of ethics in multiple disciplines, in order to find some tools for more extensive ethical considerations especially in multidisciplinary environments. This kind of a toolbox would bring more transparency into multidisciplinary research.

Turtiainen and Östman have started developing the ethical toolbox for online research already in their earlier publications (see f. ex. Turtiainen & Östman 2013; Östman & Turtiainen 2016; Östman, Turtiainen & Vaahensalo 2017). The current phase is taking the research of the research ethics into more analytical level.

2 Current research

When we are discussing such a field of research as Digital Humanities, it is quite clear than *online specific research ethics* (Östman & Turtiainen 2016; Östman, Turtiainen & Vaahensalo 2017) plays

[1] Association of Internet Researchers. AoIR has published an extensive set of ethical guidelines for online research in 2002 and 2012.

on especially significant role in it. Research projects often concentrate on one source or topic[2] with a multidisciplinary take: the understandings of research ethics may fundamentally vary even inside the same research community. Different ethical focal points and varying understandings could be a multidisciplinary resource, but it is essential to recognize and pay attention to the varying disciplinary backgrounds as well as the online specific research contexts. Only by taking these matters into consideration, we are able to create some functional ethical guidelines for multidisciplinary online-oriented research.

The Inquiries in CM24

On the basis of the two rounds of ethical inquiry within the CM24 project, the researchers seemed to consider most focal such ethical matters as anonymization, dependence on corporations, co-operation with other researchers and preserving the data. By the answers ethical views seemed to

a) individually constructed: the topic of research, methods, data plus the personal view to what might be significant
b) based on one's education and discipline tradition
c) raised from the topics and themes the researcher had come in touch with during the CM24 project (and in similar multidisciplinary situations earlier)

One thing seemingly happening with current trend of *big data* usage, is that even individually produced online material is seen as mass; faceless, impersonalized data, available to anyone and everyone. This is an ethical discussion which was already on in the early 2000's (see f. ex. Östman 2007, 2008; Turtiainen & Östman 2009) when researchers turned their interest in online material for the first time. It was not then, and it is not now, ethically durable research, to consider the private life- and everyday -based contents of individual people as 'take and run' -data. However, this seems to be happening again, especially in disciplines where ethics has mostly focused on copyrights and maybe corporal and co-operational relationships. (In the *CM24* for example information science seems to be one of the disciplines where intimate data is used as faceless mass.) Then again, a historian among the project argues in their answer, that already choosing an online discussion as an object to research is an ethical choice, "shaping what we can and should count in into the actual research".

Neither one of above-mentioned ethical views is faulty. However, it might be difficult for these two researchers to find a common understanding about ethics, in for example writing a paper together. A multifaceted, generalized collection of guidelines for multidisciplinary research would probably be of help.

Digital Humanities Journals and Publications

To explore ethics in digital humanities, we needed a diverse selection of publications to represent research in Digital Humanities. Nine different digital humanities journals were chosen for analysis,

[2] *Citizen Mindscapes* (CM24) is this kind of a project: the subject and the source are provided by *Suomi24* online discussion forum, which researchers from multiple disciplines approach from multiple viewpoints, relevant to each one's discipline.

based on the listing made by Berkeley University.[3] The focus in these journals varies from pedagogy to literary studies. However, they all are digital humanities oriented. The longest-running journal on the list[4] has been published since 1986 and the most recent journals[5] have been released for the first time in 2016. The journals therefore cover the relatively long-term history of digital humanities and a wide range of multi- and interdisciplinary topics.

In the journals and in the articles published in them, research ethics is clearly in the side, even though it is not entirely ignored. In the publications, research ethics is largely taken into account in the form of source criticism. Big data, digital technology and copyright issues related to research materials and multidisciplinary cooperation are the most common examples of research ethical considerations. Databases, text digitization and web archives are also discussed in the publications. These examples show that research ethics also affect digital humanities, but in practice, research ethics are relatively scarce in publications.

Publications of the CM project were also examined, including some of our own articles. Except for one research ethics oriented article (Östman & Turtiainen 2016) most of the publications have a historical point of view (Suominen 2016; Suominen & Sivula 2016; Saarikoski 2017; Vaahensalo 2017). For this reason, research ethics is reflected mainly in the form of source criticism and transparency. Ethics in these articles is not discussed in more length than in most of the examined digital humanities publications.

Also in this area, a multifaceted, generalized collection of guidelines for multidisciplinary research would probably be of benefit: it would be essentially significant to increase the transparency in research reporting, especially in Digital Humanities, which is complicated and multifaceted of disciplinary nature. Therefore more thorough reporting of ethical matters would increase the transparency of the nature of Digital Humanities in itself.

The Ethics Committee

The Ethics committee of the University of Turku[6] follows the development in the field of research ethics both internationally and nationally. The mission of the committee is to maintain a discussion on research ethics, enhance the realisation of ethical research education and give advice on issues related to research ethics. At the moment its main purpose is to assess and give comments on the research ethics of non-medical research that involves human beings as research subjects and can cause either direct or indirect harm to the participants.

The law about protecting personal info of private citizens appears to be a significant aspect of research ethics. Turtiainen (member of the committee) states that, at the current point, one of the main concerns seems to be poor data protection. The registers constructed of the informant base are often neglected among the humanities, whereas such disciplines as psychology and wellfare research approximately consider them on the regular basis. Then again, the other disciplines do

[3] Digital Humanities at Berkeley: http://digitalhumanities.berkeley.edu/resources/digital-humanities-journals
[4] Digital Scholarship in the Humanities: https://academic.oup.com/dsh
[5] DHCommons Journal: http://dhcommons.org/journal/2016, Digital Literary Studies: http://dhcommons.org/journal/2016 and Journal of Cultural Analytics: http://culturalanalytics.org/
[6] The Ethics committee of the University of Turku: https://www.utu.fi/fi/Tutkimus/eettisyys/eettinen-toimikunta/Sivut/home.aspx

not necessarily consider other aspects of vulnerability so deeply as the (especially culture/tradition-oriented) humanists seem to do.

Our aim is to analyse the memos of the committee, concerning the *research requests which have not been approved* (not the actual research plans, since those are classified information). Our interest focuses in arguments that have caused the rejection. It would be an interesting viewpoint to compare the rejected requests for comments from the ethics committee to the results of ethical inquiries within the CM24 project and the outline of research ethics in digital humanities journals and publications.

3 Where do you go now…

According to our current study, it seems that the position of research ethics in Digital Humanities and, more widely, in multidisciplinary research, is somewhat two-fold:

a) for example in the Digital Humanities Program of the Academy of Finland, the significance of ethics is strongly emphasized and the research projects among the program are being encouraged to increase their ethical discussions and the transparency of those. The discourse about and the interest in developing online-oriented research ethics seems to be growing and suggesting that 'something should be done'; the ethical matters should be present in the research projects in a more extensive way.
b) however, it seems that in practice the position of research ethics has not changed much within the last 10 years or so, despite the fact that the digital research environments of the humanities have become more and more multidisciplinary, which leads to multiple understandings about ethics even within individual research projects. Yet, the ethics in research reports is not discussed in more length / depth than earlier. Even in Digital Humanities -oriented journals, ethics is mostly present in a paragraph or two, repeating a few similar concerns in a way which at times seems almost 'automatic'; that is, as if the ethical discussion would have been added 'on the surface' hastily, because it is required from the outside.

This is an interesting situation. There is a possibility that researchers are not taking seriously the significance of ethical focal points in their research. This is, however, an argument that we would not wish to make. We consider it more likely that in the ever-changing digital research environment, the researches lack multidisciplinary tools for analyzing and discussing ethical matters in the depth that is needed. By examining the current situation extensively, our study is aiming at finding the focal ethical matters in multidisciplinary research environments, and at constructing at least a basic toolbox for Digital Humanities research ethical discussions.

Sources and Literature

Inquiries made by Östman, Turtiainen and Vaahensalo with the researchers the *Citizen Mindscapes 24* project. Two rounds in 2016–2017.

Digital Humanities (DigiHum). Academy Programme 2016–2019. Programme memorandum. Helsinki: Academy of Finland.

Digital Humanities journals listed by *Digital Humanities at Berkeley*. http://digitalhumanities.berkeley.edu/resources/digital-humanities-journals

Markham, Annette & Buchanan, Elizabeth 2012: *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. https://aoir.org/reports/ethics2.pdf.

Saarikoski, Petri: "Ojennat kätesi verkkoon ja joku tarttuu siihen". Kokemuksia ja muistoja kotimaisen BBS-harrastuksen valtakaudelta. *Tekniikan Waiheita* 2/2017.

Suominen, Jaakko (2016): "Helposti ja halvalla? Nettikyselyt kyselyaineiston kokoamisessa." In: Korkiakangas, Pirjo, Olsson, Pia, Ruotsala, Helena, Åström, Anna-Maria (eds.*): Kirjoittamalla kerrotut – kansatieteelliset kyselyt tiedon lähteinä*. Ethnos-toimite 19. Ethnos ry., Helsinki, 103–152. [Easy and Cheap? Online surveys in cultural studies.]

Suominen, Jaakko & Sivula, Anna (2016): "Digisyntyisten ilmiöiden historiantutkimus." In Elo, Kimmo (ed.): *Digitaalinen humanismi ja historiatieteet*. Historia Mirabilis 12. Turun Historiallinen Yhdistys, Turku, 96–130. [Historical Research of Born Digital Phenomena.]

Turtiainen, Riikka & Östman, Sari 2013: Verkkotutkimuksen eettiset haasteet: Armi ja anoreksia. In: Laaksonen, Salla-Maaria et. al. (eds.): *Otteita verkosta. Verkon ja sosiaalisen median tutkimusmenetelmät*. Tampere: Vastapaino. pp. 49–67.

– 2009: "Tavistaidetta ja verkkoviihdettä – omaehtoisten verkkosisältöjen tutkimusetiikkaa." Teoksessa Grahn, Maarit ja Häyrynen, Maunu (toim.) 2009: *Kulttuurituotanto – Kehykset, käytäntö ja prosessit*. Tietolipas 230. SKS, Helsinki. 2009. s. 336–358.

Vaahensalo, Elina: Kaikenkattavista portaaleista anarkistiseen sananvapauteen – Suomalaisten verkkokeskustelufoorumien vuosikymmenet. *Tekniikan Waiheita* 2/2017.

Östman, Sari 2007: "Nettiksistä blogeihin: Päiväkirjat verkossa." *Tekniikan Waiheita 2/2007*. Tekniikan historian seura ry. Helsinki. 37–57.

Östman, Sari 2008: "Elämäjulkaiseminen – omaelämäkerrallisten traditioiden kuopus." *Elore, vol. 15-2/2008*. Suomen Kansantietouden Tutkijain Seura. http://www.elore.fi./arkisto/2_08/ost2_08.pdf.

Östman, Sari & Turtiainen, Riikka 2016: From Research Ethics to Researching Ethics in an Online Specific Context. In *Media and Communication*, vol. 4. iss. 4. pp. 66–74. http://www.cogitatiopress.com/ojs/index.php/mediaandcommunication/article/view/571.

Östman, Sari, Riikka Turtiainen & Elina Vaahensalo 2017: From Online Research Ethics to Researching Online Ethics. Poster. *Digital Humanities in the Nordic Countries 2017 Conference.*

Copyright exceptions or licensing : how can a library acquire a digital game?

By Dr. Olivier Charbonneau, Concordia University

> Updated abstract, accepted for the DHN 2018 Digital Humanities in the Nordic Countries, 3rd Conference, 7–9 March 2018, Helsinki
>
> This abstract is available on the author's blog at this address: http://www.culturelibre.ca/dhnc-2018/
>
> The dataset used in the author's doctoral work, which is referred to in this abstract, will be made available on the Internet as soon as the final doctoral procedures are filed by the *Université de Montréal*

ABSTRACT

Copyright, caught in a digital maelstrom (Trudel, 1997) of perpetual reforms and shifting commercial practices, exacerbates tensions between cultural stakeholders. On the one hand, copyright seems to be drowned in Canada and the USA by the role reserved to copyright exceptions (Crews, 2015) by parliaments and the courts. On the other, institutions, such as libraries, are keen to navigate digital environments by allocating their acquisitions budgets to digital works (Farb, 2006; Waller, Bird, 2006). Nordic countries have explored new institutional arrangements, such as extended licensing which have proved useful for orphened works (Rosén, 2012). How can markets, social systems and institutions emerge or interact if we are not able to resolve this tension?

Beyond the paradigm shifts brought by digital technologies or globalization, one must recognize the conceptual paradox surrounding digital copyrighted works (Elkin-Koren, Salzberger, 2013). In economic terms (Maackay), they behave naturally as public goods, while copyright attempts to restore their rivalrousness and excludability. Within this paradox lies tension, between the aggregate social wealth (Frishmann, 2012; Yoo, 2007) spread by a work and its commoditized value, between network effects (Benkler, 2006) and reserved rights.

In this paper, I will summarize the findings of my doctoral research project and apply them to the case of digital games in libraries.

The goal of my doctoral work was to ascertain the role of libraries (Hirtke, Hudson & Kenyon, 2009; Larivière, 1989; Gordon, 1982) in the markets and social systems (Luhmann, 2004) of digital copyrightable works. Ancillary goals included exploring the "border" between licensing and exceptions in the context of heritage institutions as well as building a new method for capturing the complexity of markets and social systems that stem from digital protected works. To accomplish these goals, I analysed a dataset comprising the terms and conditions of licenses held by academic libraries in Québec. I show that the terms of these licences overlap with copyright exceptions, highlighting how Libraries express their social mission in two normative contexts (Belley, 1996), positive law (copyright exceptions) and private ordering (licensing). This overlap is both necessary yet poorly understood – they are not two competing institutional arrangements but the same image reflected in two distinct social settings. It also provides a road-map for right-holders of how to make digital content available through libraries.

The study also points to the rising importance of automation and computerization in the provisioning of licences (Garnett, 2006; Mountain, 2003; Gillette, Radin, 2000) in the digital world. Metadata describing the terms of a copyright licence (Maurel, 2007) are increasingly represented in computer models and leveraged to mobilize digital corpus for the benefit of a community. Whereas the print world was driven by assumptions and physical limits to using copyrighted works, the digital environment introduces new data points for interactions which were previously hidden from scrutiny. The future lies not in optimizing transaction costs but in crafting elegant institutional arrangements through licensing.

If libraries exist to capture some left-over value in the utility curve (again: Frishmann, 2012; Yoo, 2007) of our cultural, informational or knowledge markets, the current role they play in copyright need not change in the digital environment. What does change, however, is hermeneutics: how we attribute value to digital copyrighted works and how we study society's use of them.

We will transpose the results of this study to the case of digital games. Québec and Montréal are currently hotbeds for both independent (indie) and major (AAA) video game studios. Despite this, a market failure currently exists due to the absence of flexible licensing mechanisms to make indie games available through libraries. This part of the study was funded with the generous support from the Knight Foundation in the USA and conducted at the Technoculture Art & Games (TAG) research cluster of the Milieux Institute for arts, culture and technology at Concordia University in Montréal, Canada.

Most notably, Libraries must still rely on physical copies of media to constitute their collections of digital games. This is mostly due to restrictive licensing from all major game platforms (iTunes, Google Play, Steam, etc.), which only allow individuals to transact. In that sense, platforms or marketplaces providing access to digital games – and most other digital works for that matter – are closed to libraries. This is in stark contrast to markets for physical books or that of physical media such as disks – because exclusionary practices are eschewed or because of copyright exceptions (such as the case of the first sale doctrine in the United-States).

Using licensing to ward off libraries from digital markets will be problematic in the medium to long term. In fact, because digital works are simply unaccessible by libraries, it is unclear how copyright exceptions as well as extended licensing, as implemented in Nordic Countries, could solve this issue. No amount of legislative reform or litigation could have an impact on this problem in the short term. We hope that new institutional arrangements, achieved through private ordering or licensing, could point the way forward.

For example, to solve this issue for e-books, public libraries in Québec created a new collecting society to foster the emergence of a new institutional arrangement with publishers. Using the concept of the commons (Benkler, 2017), our research team at Concordia University is seeking a similar path for born-digital games, prototyping a new institutional arrangement using blockchain technologies. Our goal is to allow for the drafting of smart contracts, fostering micro-transactions on an open ledger as well as using asynchronous encryption to open this digital marketplace.

In turn, we hope that this model will allow libraries to play their role for society and allow society to reap the positive externalities of its libraries' mission. Libraries may even be able to dent the market dominance of existing digital platforms.

References

- Benkler, Yochai. "Open-Access and Information Commons." In *The Oxford Handbook of Law and Economics: Volume 2: Private and Commercial Law*, edited by Francesco Parisi, 257–79. Oxford University Press, 2017.
- Belley, J. G. 1996. "Le contrat comme phénomène d'internormativité." In *Le droit soluble : contributions québécoises à l'étude de l'internormativité*, edited by J. G. Belley, 195–232. Paris: L.G.D.J.
- Belley, Jean-Guy. 1992. "Les Transformations d'un Ordre Juridique Privé. Les Contrats d'approvisionnement à l'ère de La Cybernétique et de La Gestion Stratégique." *Les Cahiers de Droit* 33 (1): 21–70. https://doi.org/10.7202/043126ar.
- Crews, Kenneth D. 2015. "Study on Copyright Limitations and Exceptions for Libraries and Archives: Updated and Revised." SCCR/30/3. Standing Committee on Copyright and Related Rights. Geneva: World Intellectual Property Organisation. http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=306216.
- Elkin-Koren, Niva, and Eli M. Salzberger. 2013. *The Law and Economics of Intellectual Property in the Digital Age : The Limits of Analysis*. Book, Whole. Abingdon, Oxon England]; New York: Routledge.
- Farb, Sharon. 2006. "Libraries, Licensing and the Challenge of Stewardship." *First Monday* 11 (7). http://firstmonday.org/ojs/index.php/fm/article/view/1364.
- Frischmann, Brett M. 2012. *Infrastructure : The Social Value of Shared Resources*. Book, Whole. New York: Oxford University Press.
- Garnett, Nic. 2006. "Automated Rights Management Systems and Copyright Limitations and Exceptions." WIPO. http://www.wipo.int/edocs/mdocs/copyright/en/sccr_14/sccr_14_5.pdf.
- Gillette, Clayton P., and Margaret Jane Radin. 2000. "Interpretation and Standardiziation in Electronic Sales Contracts XML and the Legal Foundations for Electric Commerce Online Standardization and the Integration of Text and Machine The Robert L. Levine Distinguished Lecture Series." *Southern Methodist University Law Review* 53: 1431–46.
- Gordon, Wendy J. 1982. "Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors." *Columbia Law Review* 82: 1600–1657.
- Hirtle, Peter B., Emily Hudson, and AndrewT Kenyon. 2009. *Copyright & Cultural Institutions : Guidelines for Digitization for U.S. Libraries, Archives, & Museums*. Monograph. Ithaca (New York): Cornell University Library.
- Hudson, Emily Burrell, Robert. 2011. "Abandonment, Copyright and Orphaned Works: What Does It Mean to Take the Proprietary Nature of Intellectual Property Rights Seriously." *Melbourne University Law Review*35: 971–1004.
- Larivière, Jules. 1989. "Les Exceptions Applicables Aux Bibliothèques et Aux Centres de Documentation En Matière de Droit d'auteur." *Documentation et Bibliothèques* 35 (4): 135–42.
- Luhmann, Niklas, Klaus A. Ziegert, and Fatima Kastner. 2004. *Law as a Social System*. Oxford Socio-Legal Studies, Book, Whole. Oxford; New York: Oxford University Press.
- Mackaay, Ejan. 2008. "The Economics of Intellectual Property Rights in Civil Law Systems." In *Economic Analysis of Law – A European Perspective*, edited by Aristides N. Hatzis, 1–23. Cheltenham, UK: Edward Elgar.
- ———. 2013. *Law And Economics For Civil Law Systems*. Northampton, MA: Edward Elgar.
- Mackaay, Ejan, and Stéphane Rousseau. 2008. *Analyse Économique Du Droit*. Vol. 2e éd. Méthodes Du Droit, Book, Whole. Montréal; Paris: Éditions Thémis; Dalloz.
- Maurel, Lionel. 2007. "Panorama Des Systèmes de Métadonnées Juridiques et de Leurs Applications En Bibliothèque Numérique." *Les Cahiers de Propriété Intellectuelle* 19 (1): 241–76.
- Mountain, Darryl. 2003. "XML E-Contracts: Documents That Describe Themselves." *Int'l J.L. & Info. Tech.* 11: 274–85.
- Rosén, Jan. 2012. "La Diffusion En Ligne et Le Régime de Licence Collective Étendue (« ECL ») Des Pays Nordiques – Les Œuvres Orphelines Comme Précédent." *Les Cahiers de Propriété Intellectuelle* 24 (2): 321–46. http://cpi.robic.ca/Cahiers/CPI%2024-2/CPI%2024-2%20mai%202012.pdf.

- Trudel, Pierre. 1997. *Droit Du Cyberespace*. Monograph. Montréal: Les Éditions Thémis.
- Waller, Andrew, and Gwen Bird. 2006. "'We Own It': Dealing with « Perpetual Access in Big Deals." *Serials Librarian* 50 (1–2): 179–96. http://0-search.ebscohost.com.mercury.concordia.ca/login.aspx?direct=true&db=a9h&AN=22526189&site=ehost-live&scope=site.
- Yoo, Christopher S. 2007. "Copyright and Public Good Economics: A Misunderstood Relation." *University of Pennsylvania Law Review*155 (3): 635–715. https://doi.org/10.2307/40041335.

Ce contenu a été mis à jour le 5 février 2018 à 12 h 15 min.

Poster presentation abstract
Digital Humanities in the Nordic Countries 2018, Helsinki, Finland

Elisabeth Stubb, PhD
Svenska litteratursällskapet i Finland
Albert Edelfelts brev
elisabeth.stubb@sls.fi

Shearing letters and art as digital cultural heritage, co-operation and basic research

Albert Edelfelts brev (edelfelt.fi) is a web publication developed at the Society of Swedish Literature in Finland. In co-operation with the Finnish National Gallery, we publish letters of the Finnish artist Albert Edelfelt (1854–1905) combined with pictures of his artworks. *Albert Edelfelts brev* received in 2016 the State Award for dissemination of information. In 2017 Albert Edelfelt's letter collections were adopted as a Cultural Heritage in the National Register of Unesco's Memory of the World (http://www.maailmanmuisti.fi/p/suomen-kansallinen-maailman-muisti.html). The co-operation between institutions and basic research of the material has enabled a unique reconstruction of Edelfelt's artistry and his time, for the service of researchers and other users. I will present how we have done it and how we plan to further develop the website.

The nature of the publication project encompasses two communities that have emerged in digital arts and humanities research. The Society of Swedish Literature draws its tradition of web publications mainly from Digital Humanities, which has appeared preoccupied with transforming the traditions of text-based humanities computing, drawn from scholarly practice. The material in *Albert Edelfelts brev* is, on the other hand, also connected to Digital Heritage, which has drawn more from theories and practices in the digital representation of material culture. As a third element in developing the project emerges the users, and how they in a relevant way are able to use the material. (Benardou, Champion, Dallas, Hughes, 2018).

The website *Albert Edelfelts brev* launched in September 2014, with a sample of Edelfelt's letters and paintings. Our intention is to publish all the letters Albert Edelfelt wrote to his mother Alexandra (1833–1901). The collection consists of 1 310 letters, that range over 30 years and cover most of Edelfelt's adult life. The letters are in the care of the Society of Swedish Literature in Finland. We also have to our disposal close to 7 000 pictures of Edelfelt's paintings and sketches in the care of the Finnish National Gallery.

In a digital context, the volume of the material at hand is manageable. However, for researchers who think that they might have use of the material, but are unsure of exactly where or what to look for, it might be labour intensive to go through all the letters and pictures. We have combined professional expertise and basic research of the material with digital solutions to make it as easy as possible to take part of what the content can offer.

As editor of the web publication, I spend a considerable part of my work on basic research in identifying people, and pinpointing paintings and places that Edelfelt mentions in his letters.

By linking the content of a letter to *artworks*, *persons*, *places* and *subjects/reference words* users can easily navigate in the material. Each letter, artwork and person has a page of its own. Even places and subjects are searchable and listed.

The letters are available as facsimile pictures of the handwritten pages. Each letter has a permanent web resource identifier (URN:NBN). In order to make it easier for users to decide if a letter is of interest, we have tagged subjects using reference words from ALLÄRS (common thesaurus in Swedish). We have also written abstracts of the content, divided them into separate "events" and tagged mentioned artworks, people and places to these events.

Each artwork of Edelfelt has a page of its own. Here, users find a picture of the artwork (if available) and earlier sketches of the artwork (if available). By looking at the pictures, they can see how the working process of the painting has developed. Users can also follow the process through Edelfelt's writings in his letters. All the events from the letter abstracts that are tagged to the specific artwork are listed in chronological order on the artwork-page.

Persons tagged in the letter abstracts also have pages of their own. On a person-page, users find basic facts and links to other webpages with information about the person. Any events from the letter abstracts mentioning the person are listed as well. In other words, through a one-click-solution users can find an overview on everything Edelfelt's letters have to say about a specific person. Tagging persons to events, has also made it possible to build graphs of a person's social network; based on how many times other persons are tagged to the same events as the specific person. There is a link to these graphs on every person-page.

Apart from researchers who have a direct interest in the material, we have also wanted to open up the cultural heritage to a broader public and group of users. Each month the editorial staff writes a blog-post on *SLS-bloggen* (http://www.sls.fi/sv/blogg/tag/edelfelt-0). *Albert Edelfelts brev* also has a profile on Facebook (https://www.facebook.com/albertedelfeltsbrev/) where we post excerpts of letters on the same date as Edelfelt wrote the original letter. This way, we hope to give the public an insight in the life of Edelfelt and the material, and involve them in the progress of the project.

The web publication has open access. The mix of different sources and the co-operation with other heritage institutions has led to a mix of licenses for how users can copy and redistribute the published material. The Finnish National Gallery (FNG) owns copyright on its pictures in the publication and users need permission from FNG to copy and redistribute their material. The artwork-pages contain descriptions of the paintings written by the art historian Bertel Hintze, who published a catalogue of Edelfelt's art in 1942. These texts are licensed with a Creative Commons Attribution-NoDerivs 4.0 Generic (CC BY-ND 4.0). Edelfelt's letters, as well as the texts and metadata produced by the editorial staff at the Society of Swedish Literature in Finland, have a Creative Commons CC0 1.0 Universal-license. Data with Creative Commons-license is also freely available as open data through a REST API (http://edelfelt.sls.fi/apiinfo/).

In the future, we would like to find a common practice for the user rights; if possible, that all the material would have the same license. We intend to invite other institutions with artworks

of Edelfelt to co-operate, offering the same kind of partnership as the web publication has with the Finnish National Gallery. Thus, we are striving to a complete as possible site with the artworks of Edelfelt.

Albert Edelfelt is of national interest and his letters, which he mostly wrote during his stays abroad, contain information of also international interest. Therefore, we plan to offer the metadata and some of the source material in Finnish and, hopefully, English translations. So far, the letters are only available as facsimile. The development of transcription programs for handwritten texts has made it probable that we in the future include transcriptions of the letters in the web publication. Linguists especially have an interest in getting a searchable letter transcription for their researches, and the transcriptions would also be helpful for users who might have problem reading the handwritten text.

References:

Albert Edelfelts brev. Elektronisk brev- och konstutgåva, utg. Maria Vainio-Kurtakko & Henrika Tandefelt & Elisabeth Stubb, Svenska litteratursällskapet i Finland, 2014–, http://edelfelt.fi/

Benardou, Agiatis & Eric Champion & Costis Dallas & Lorna M. Hughes, "Introduction: a critique of digital practices and research infrastructures", *Cultural Heritage Infrastructures in Digital Humanities*, Routledge 2018.

Abstract/short presentation

Revisiting the authorship of Henry VIII's *Assertio septem sacramentorum* through computational authorship attribution

Presenters: Marjo Kaartinen, Aleksi Vesanto, Anni Hella

One of the great mysteries of Tudor history through centuries has been the authorship of Henry VIII's famous treatise *Assertio septem sacramentorum adversus Martinum Lutherum* (1521). The question of its authorship intrigued the contemporaries already in the 1520s. With *Assertio*, Henry VIII gained from the Pope the title *Defender of the Faith* which the British monarchs still use. Because of the exceptional importance of the text, the question of its authorship is not irrelevant in the study of history. This paper contributes to the conference theme *History*.

For various reasons and motivations each of their own, many doubted the king's authorship. The discussion has continued to the present day. Many possible authors have been named, Thomas More foremost among them. The Academy of Finland funded consortium Profiling Premodern Authors (PROPREAU) aims at giving new light to the question of Assertio's authorship. We have begun with examining the possibility of More's authorship, and the first tentative results will be revealed at the conference.

Our consortium puts effort into developing more efficient machine learning methods for authorship attribution in cases where large training corpora are not available. This paper will present the latest discoveries in the development of such tools. These include simpler, linear classifiers and more complex neural networks. Support vector machine, a linear classifier produces fairly accurate early results, though they alone are most likely not enough. Due to this, we also use a convolutional neural network to reinforce the results to make more solid interpretations.

Select Bibliography:
Betteridge, Thomas: *Writing Faith and Telling Tales: Literature, Politics, and Religion in the Work of Thomas More*. University of Notre Dame Press 2013.
Brown, J. Mainwaring: Henry VIII.'s Book, "Assertio Septem Sacramentorum," and the Royal Title of "Defender of the Faith". *Transactions of the Royal Historical Society* 1880, 243–261.
Nitti, Silvana: *Auctoritas: l'Assertio di Enrico VIII contro Lutero.* Studi e testi del Rinascimento europeo. Edizioni di storia e letteratura 2005.

ARKWORK: Archaeological practices and knowledge in the digital environment

Poster abstract

Archaeology and material cultural heritage have often enjoyed a particular status as a form of heritage that has captured the public imagination. As researchers from many backgrounds have discussed, it has become the locus for the expression and negotiation of European, local, regional, national and intra-national cultural identities, for public policy regarding the preservation and management of cultural resources, and for societal value in the context of education, tourism, leisure and well-being. The material presence of objects and structures in European cities and landscapes, the range of archaeological collections in museums around the world, the monumentality of the major archaeological sites, and the popular and non-professional interest in the material past are only a few of the reasons why archaeology has become a linchpin in the discussions on how emerging digital technologies and digitization can be leveraged for societal benefit. However, at the time when nations and the European community are making considerable investments in creating technologies, infrastructures and standards for digitization, preservation and dissemination of archaeological knowledge, critical understanding of the means and practices of knowledge production in and about archaeology from complementary disciplinary perspectives and across European countries remains fragmentary, and in urgent need of concertation.

In contrast to the rapid development of digital infrastructures and tools for archaeological work, relatively little is known about how digital information, tools and infrastructures are used by archaeologists and other users and producers of archaeological information such as archaeological and museum volunteers, avocational hobbyists, and others. Digital technologies (infrastructures, methods and resources) are reconfiguring aspects of archaeology across and beyond the lifecycle (i.e., also "in the wild"), from archaeological data capture in fieldwork to scholarly publication and community access/entanglement.Both archaeologists and researchers in other fields, from disciplines such as museum studies, ethnology, anthropology, information studies and science and technology studies have conducted research on the topic but so far, their efforts have tended to be somewhat fragmented and anecdotal. This is surprising, as the need of better understanding of archaeological practices and knowledge work has been identified for many years as a major impediment to realizing the potential of infrastructural and tools-related developments in archaeology. The shifts in archaeological practice, and in how digital technology is used for archaeological purposes, calls for a radically transdisciplinary (if not interdisciplinary) approach that brings together perspectives from reflexive, theoretically and methodologically-aware archaeology, information research, and sociological, anthropological and organizational studies of practice.

This poster presents the COST Action “Archaeological practices and knowledge work in the digital environment” (http://www.cost.eu/COST_Actions/ca/CA15201 - ARKWORK), an EU-funded network which brings together researchers, practitioners, and research projects studying archaeological practices, knowledge production and use, social impact and industrial potential of archaeological knowledge to present and highlight the on-going work on the topic around Europe.

ARKWORK (https://www.arkwork.eu/) consists of four Working Groups (WGs), with a common objective to discuss and practice the possibilities for applying the understanding of archaeological knowledge production to tackle on-going societal challenges and the development of appropriate management/leadership structures for archaeological heritage. The individual WGs have the following specific but complementary themes and objectives:

WG1 - Archaeological fieldwork
Objectives: To bring together and develop the international transdisciplinary state-of-the-art of the current multidisciplinary research on archaeological fieldwork. How archaeologists are conducting fieldwork and documenting their work and findings in different countries and contexts and how this knowledge can be used to make contributions to developing fieldwork practices and the use and usability of archaeological documentation by the different stakeholder groups in the society.

WG2 - Knowledge production and archaeological collections
Objectives: To integrate and push forward the current state-of-the-art in understanding and facilitating the use and curation of (museum) collections and repositories of archaeological data for knowledge production in the society.

WG3 - Archaeological knowledge production and global communities
Objectives: To bring together and develop the current state-of-the-art on the global communities (including indigenous communities, amateurs, neo-paganism movement, geographical and ideological identity networks and etc.) as producers and users in archaeological knowledge production e.g. in terms of highlighting community needs, approaches to communication of archaeological heritage, crowdsourcing and volunteer participation.

WG4 - Archaeological scholarship
Objectives: To integrate and push forward the current state-of-the-art in study of archaeological scholarship including academic, professional and citizen science based scientific and scholarly work.

In our poster we outline each of the working groups and provide a clear overview of the purposes and aspirations of the COST Action Network ARKWORK.

“Everlasting Runes”: A Research Platform and Linked Data Service for Runic Research

“Everlasting Runes” (Swedish: “Evighetsrunor”) is a three-year collaboration between the Swedish National Heritage Board and Uppsala University, with funding provided by the Bank of Sweden Tercentenary Foundation (Riksbankens jubileumsfond) and the Royal Swedish Academy of Letters (Kungliga Vitterhetsakademien). The project combines philology, archaeology, linguistics, and information systems, and is comprised of several research, digitisation, and digital development components.

Chief among these is to release both existing data and newly digitised material onto the web as openly-licensed linked data, in conjunction with the development of a web-based research platform for runic researchers, with the aim of drawing together disparate structured digital runic resources into a single interface, conveniently accessible for both human and machine agents. As part of the platform’s development, the corpus of Scandinavian runic inscriptions in Uppsala University’s *Runic Text Database* (*Samnordisk runtextdatabas*) will be restructured and marked up for use on the web, and linked against their entries in the previously digitised standard corpus work (*Sveriges runinskrifter*). In addition, photographic archives of runic inscriptions from the 19th- and 20th centuries from both the Swedish National Heritage Board archives and Uppsala University library will be digitised, alongside other hitherto inaccessible archive material. Information from a number of existing relevant open data resources will also be linked in to the platform.

As a collaboration between a university and a state heritage agency with a small research community as its primary target audience, the project must bridge the gap between the different needs and abilities of these stakeholders, as well as resolve issues of long-term maintenance and stability which have previously proved problematic for some of the source datasets in question. The project must also contend with the differing theoretical backgrounds and research needs of the research communities that it intends to benefit – primarily runologists and archaeologists, but also philologists, corpus linguists, and art historians – by structuring and representing the data in a sound manner that is both accessible and useful to these disciplines but also generalised enough to take account of the material’s diversity and ensure interoperability with other cultural heritage data sets on the broader semantic web. It is hoped that the resulting research- and data platforms will combine the strengths of both the National Heritage Board and Uppsala university to produce a rich, actively-maintained scholarly resource.

This paper will present the background and aims of the project within the context of runic research, as well as the various datasets that will be linked together in the linked data research platform (with its corresponding web interface) with particular focus on the data structures in question, the philological markup of the corpus of inscriptions, and requirements gathering.

Topics: historical studies, audio / video / multimedia, digital resources – publication and discovery, history and theory of digital humanities

Keywords: television, reporting, environment, availability, comparative studies

Digital humanities and environmental reporting in television during the Cold War
Methodological issues of exploring materials of the Estonian, Finnish, Swedish, Danish, and British broadcasting companies

Simo Laakkonen

University of Turku, Degree Programme on Cultural Production and Landscape Studies;
simo.laakkonen@utu.fi

Environmental history studies have relied on traditional historical archival and other related source materials so far. Despite the increasing availability of new digitized materials studies in this field have not reacted to these emerging opportunities in any particular way. The aim of the proposed paper is to discuss possibilities and limitations that are embodied in the new digitized source materials in different European countries. The proposed paper is an outcome of a research project that explores the early days of television prior to the Earth Day in 1970 and frame this exploration from an environmental perspective. The focus of the project is reporting of environmental pollution and protection during the Cold War. I had originally planned a comparative study of environmental broadcasting in the Baltic Sea Region including western democracies and a Communist state. In order to realize this study the quantity and quality of related digitized and non-digitized source materials provided by the national broadcasting companies of Estonia (ETV), Finland (YLE), Sweden (SVT), Denmark (DR), and United Kingdom (BBC) were examined. However, realising a comparative project proved impossible. Only a few programs on environmental topics from Soviet times were found in the archives of Estonian television (ETV) that were otherwise superb and easy to use. The Swedish television (SVT) archives, on the other hand, censored its qualitative bodies of data by denying access to printed content reports. Danish television (DR) was very helpful and provided requested content reports without any limitations. In addition, a visit was made to the renowned BBC television archives in Reading, England. But the BBC's archival material from the early days was of poor quality, and their content reports had not been digitized. The data contained in the archives managed by Finnish Public Broadcasting (YLE) and by Danish television (DR) proved, thus, to be a superb source for this research. The main outcome of this international comparative study of available sources is that the quantity and quality of available materials varies greatly, even in a surprising way between the examined countries that belonged to different political spheres (Warsaw Pact, neutral, NATO) during the Cold War. Hence I am not suggesting abandoning the virtues of old-school historical methodology. Rather I propose combining the best of the old-school approaches and the potential of the new school to look for a wider array of source materials, alternative methods for searching for and processing information, and new channels for publishing and popularizing results.

*

Simo Laakkonen is a senior lecturer of landscape studies in the Degree Programme of Cultural Production and Landscape Studies (University of Turku, Finland), which includes three subfields: landscape studies, digital culture, and cultural heritage. Previously he has worked

as a lecturer and acting professor of environmental policy at the University of Helsinki and University of Eastern Finland. He has directed several international research projects on the environmental history of the Baltic Sea. He has published or edited several books and coedited two special issues on socioecological Baltic Sea questions published in *Ambio: A Journal of the Human Environment* published by Royal Swedish Academy of Sciences. Over the past decades he has explored the environmental history of WWII and the Cold War that fundamentally reshaped environment and societies in the Baltic Sea region. His latest publication is a co-edited volume *The Long Shadows: A Global Environmental History of the Second World War* (Corvallis: Oregon State University Press, 2017).

The Bank of Finnish Terminology in Arts and Sciences – a new form of academic collaboration and publishing

Johanna Enqvist & Tiina Onikki-Rantajääskö (University of Helsinki)

Our poster presents the multidisciplinary research infrastructure project “Bank of Finnish Terminology in Arts and Sciences (BFT)” (“The Helsinki Term Bank for the Arts and Sciences [HTB]” from 2018) as an innovative form of academic collaboration and publishing. The BFT, which was launched in 2011, aims to build a permanent and continuously updated terminological database for all fields of research in Finland. The project maintains a wiki-based website (http://tieteentermipankki.fi) which offers an open and collaborative platform for terminological work and a discussion forum available to all registered users. The BFT has been acknowledged as an innovation due to its unique combination of wiki-based technology, concentration on research terminology, and communal, open and democratic working methods in the academic context.

The BFT thus opens not only the results but the whole academic procedure where the knowledge is constantly produced, evaluated, linked, discussed and updated in an ongoing process. The BFT also provides an inclusive arena for all the interested people – students, journalists, translators and enthusiasts – to participate in the discussions relating to concepts and terms in Finnish research. Based on the knowledge and experiences accumulated during the BFT project we will reflect on the benefits, challenges, and future prospects of this innovative and globally unique approach. Furthermore, we will consider the possibilities and opportunities opening up especially in terms of digital humanities.

Content for the BFT is created by niche-sourcing, where the participation is limited to a particular group of experts in the participating subject fields. Niche-sourcing invites the research community to take responsibility for the availability of up-to-date terminology in their research fields. Sharing this responsibility among experts has its advantages: it is possible to get the best specialists involved and thus guarantee the quality and accuracy of the content. Furthermore, when shared among the group of experts the task is not too large for any individual participant. This is also a more democratic way of carrying out terminology work: there is no single gatekeeper in the field. Using the BFT for online discussions the expert groups are also able and encouraged to create their own plans of action and general guidelines concerning the terminological work in their field.

Today, the BFT includes 40 subject fields, and the database contains over 40 000 concept pages and 300 000 terms/lemmas. The amount of traffic in the BFT website has been growing steadily in terms of both visits and users, and at the moment there are over 18 000 weekly visits. However, the BFT is still in the construction phase, when it comes to its contents. In regard to that, we are aiming for extensive expansion during the next three years. Our objective is to have at least ten new subject fields included in the bank each year. As the content of the BFT increases, the usability, reliability and validity of the BFT as digital resource and data for research also improves.

The potential and future prospects for the BFT are various. The BFT as digital resource could be useful for instance in designing and developing applications, such as automatic text genre recognition or term extraction tools; or in analysing and comparing conceptualisations and discourses in and between the different fields of research; or examining the method of niche-sourcing in the academic context. There is also need for multilingual and international

cooperation in terminology work for which the BFT offers a platform. Therefore, we are constantly looking for partners focused especially on research terminology.

ArchiMob: A multidialectal corpus of Swiss German oral history interviews

Yves Scherrer, University of Helsinki
Tanja Samardžić, University of Zurich

Although dialect usage is prevalent in the German-speaking part of Switzerland, digital resources for dialectological and computational linguistic research are difficult to obtain. In this paper, we present a freely available corpus of spontaneous speech in various Swiss German dialects. It consists in transcriptions of video interviews with contemporary witnesses of the Second World War period in Switzerland. These recordings were produced by an association of Swiss historians called Archimob[1] about 20 years ago. More than 500 informants originating from all linguistic regions of Switzerland (German, French and Italian) and representing both genders, different social backgrounds, and different political views, were interviewed. Each interview is 1 to 2 hours long. In collaboration with the University of Zurich, we have selected, processed and analyzed a subset of 43 interviews in different Swiss German dialects.

The goal of this contribution is twofold. First, we describe how the documents were transcribed, segmented and aligned with the audio source and how we make the data available on specifically adapted corpus query engines. We also provide an additional layer in which each transcribed word form is associated with a normalized form. This normalization reduces the different types of variation (dialectal, speaker-specific and transcriber-specific) present in the transcriptions; the normalization language resembles standard German. We formalize normalization as a machine translation task, obtaining up to 90% of accuracy (Scherrer & Ljubešić 2016).

Second, we show through some examples how the ArchiMob resource can shed new lights on research questions from digital humanities in general and dialectology and history in particular:

- Thanks to the normalization layer, dialect differences can be identified and compared with existing dialectological knowledge.
- Using language modelling, another technique borrowed from language technology, we can compute distances between texts. These distance measures allow us to identify the dialect of unknown utterances (Zampieri et al. 2017), localize transcriber effects and obtain a generic picture of the Swiss German dialect landscape.
- Departing from the purely formal analysis of the transcriptions for dialectological purposes, we can apply methods such as collocation analysis to investigate the content of the interviews. By identifying the key concepts and events referred to in the interviews, we can assess how the different informants perceive and describe the same time period.

References

Tanja Samardžić, Yves Scherrer & Elvira Glaser (2016): *ArchiMob - A corpus of Spoken Swiss German*. Proceedings of LREC 2016, 4061-4066, Portorož, Slovenia.

Yves Scherrer & Nikola Ljubešić (2016): *Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation*. Proceedings of KONVENS 2016, 248-255, Bochum, Germany.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, Noëmi Aepli (2017): *Findings of the VarDial Evaluation Campaign 2017*. Proceedings of the VarDial 2017 Workshop, EACL, Valencia, Spain.

[1] Archimob stands for *Archives de la mobilisation*, i.e. archives of the mobilization period.

Abstract

“See me! Not my gender, race, or social class”:
Combating Stereotyping and prejudice mixing digitally manipulated experience with classroom debriefing.

INTRODUCTION

Not only does stereotyping, based on various social categories such as age, social class, ethnicity, sexuality, regional affiliation, and gender serve to simplify how we perceive and process information about individuals (Talbot 2003: 468), it also builds up expectations on how we act. If we recognise social identity as an ongoing construct, and something that is renegotiated during every meeting between humans (Crawford 1995), it is reasonable to speculate that stereotypic expectations will affect the choices we make when interacting with another individual. Thus, stereotyping may form the basis for the negotiation of social identity on the micro level. For example, research has shown that white American respondents react with hostile face expressions or tone of voice when confronted with African American faces, which is likely to elicit the same behaviour in response, but, as Bargh et al. point out (1996: 242), “because one is not aware of one's own role in provoking it, one may attribute it to the stereotyped group member (and, hence, the group)”. Language is a key element in this process. Our hypothesis is that linguistic stereotyping acts like a filter making us notice those features which we expect to find, and toning down other features. An awareness of such phenomena, and how we unknowingly may be affected by the same, is, we would argue, essential for all professions where human interaction is in focus (psychologists, teachers, social workers, health workers etc.).

RAVE (Raising Awareness through Virtual Experiencing) funded by the Swedish Research Council, aims to explore and develop innovative pedagogical methods for raising subjects’ awareness of their own linguistic stereotyping, biases and prejudices, and to systematically explore ways of testing the efficiency of these methods. The approach is the use of digital matched-guise testing techniques, thus upgrading an established method for recording stereotypical views on accents (Lambert et al 1960) with regard to applicability, and reproducible and transparent research practices. Previously matched-guise techniques could not be applied to gender studies, and even in accent studies, there was the issue of the uniqueness of each recording. However, with digital methods, it is now possible to create two versions of the same recording differing only with regard to one variable (e.g. perception gender codified in terms of pitch and timbre) in a procedure that is fully reproducible and transparent.

We are confident that there is a place for this, in our view, timely product. There can be little doubt that the zeitgeist of the 21st centuries first two decades has swung the pendulum in a direction where it has become apparent that the role of Humanities should be central. In times

when unscrupulous politicians take every chance to draw on any prejudice and stereotypical assumptions about Others, be they related to gender, ethnicity or sexuality, it is the role of the Humanities to hold up a mirror and let us see ourselves for what we are. This is precisely the aim of the RAVE project.

In line with this thinking, open access to our materials and methods is of primary importance. Here our ambition is not only to provide tested sample cases for open access use, but also to provide clear directives on how these have been produced so that new cases, based on our methods, can be created. This includes clear guidelines as to what important criteria need to be taken into account when so doing, so that our methodology is disseminated openly and in such a fashion that it becomes adaptable to new contexts.

METHOD

The RAVE method at its core relies on a treatment session where two groups of test subjects (i.e. students) each are exposed to one out of two different versions of the same scripted dialogue. The two versions differ only with respect to the perception of the gender of the characters, whereas scripted properties remain constant. In one version, for example, one participant, "Terry", may sound like a man, while in the other recording this character has been manipulated for pitch and timbre to sound like a woman. After the exposure, the subjects are presented with a survey where they are asked to respond to questions related to linguistic behaviour and character traits one of the interlocutors. The responses of the two sub-groups are then compared and followed up in a debriefing session, where issues such as stereotypical effects are discussed.

The two property-bent versions are based on a single recording, and the switch of the property (for instance, gender) is done using digital methods described below. The reason for this procedure is to minimize the number of uncontrolled variables that could affect the outcome of the experiment. It is a very difficult - if not an impossible - task to transform the identity-related aspects of a voice recording, such as gender or accent, while maintaining a "perfect" and natural voice - a voice that is opposite in the specific aspect, but equivalent in all other aspects, and doing so without changing other properties in the process or introducing artificial artifacts.

Accordingly, the RAVE method doesn't strive for perfection, but focuses on achieving a *perceived credibility* of the scripted dialogue. However, the base recording is produced with a high quality to provide the best possible conditions for the digital manipulation. For instance, the dialogue between the two speakers are recorded on separate tracks so as to keep the voices isolated.

The digital manipulation is done with the Praat software (Boersma & Weenink, 2013). Formants, range and and pitch median are manipulated for gender switching using standard offsets and are then adapted to the individual characteristics of the voices. Several versions of the manipulated dialogues are produced, and evaluated by a test group via an online survey. Based on the survey result, the one with the highest quality is selected. This manipulated dialogue needs further framing to reach a sufficient level of credibility.

The way the dialogue is framed for the specific target context, how it is packaged and introduced is of critical importance. Various kinds of techniques, for instance use of audiovisual cues, are used to distract the test subject from the "artificial feeling", as well as to enforce the desired target property. We add various kinds of distractions, both audial and visual, which lessen the listeners' focus on the current speaker, such as background voices simulating the dialogue taking place in a cafe, traffic noise, or scrambling techniques simulating, for instance, a low-quality phone or a Skype call.

On this account, the RAVE method includes a procedure to evaluate the overall (perceived) quality and credibility of a specific case setup.This evaluation is implemented by exposing a number of pre-test subjects to the packaged dialogue (in a set-up comparable to the target context). After the exposure, the pre-test subjects respond to a survey designed to measure the combined impression of aspects such as the scripted dialogue, the selected narrators, the voices, the overall set-up, the contextual framing etc.

The produced dialogues, and accompanying response surveys are turned into a single online package using the program Storyline. The single entry point to the package makes the process of collecting anonymous participant responses more fail-safe and easier to carry out.

The whole package is produced for a "bring your own device" set-up, where the participants use their own smartphones, tablets or laptops to take part in the experiment. These choices of using an online single point of entry package adapted to various kinds of devices have been made to facilitate experiment participation and recording of results. The results from the experiment is then collected by the teacher and discussed with the students at an ensuing debriefing seminar.

In the debriefing seminar after the exposure, students, organized in small groups, have an opportunity to reflect on the results from the experiment. Since any difference between the groups was the result of the participants' rating, their own reactions to the conversations, there is something very concrete and urgent to discuss. Thus, the pedagogical application for the set-up is to confront students or other participants with their own stereotypical assumptions. With the method described here, where the dialogues are identical except for the digital manipulation,

perceived differences in personality and social behaviour can only be explained as residing in the beholder.

FINDINGS

At this stage, we have conducted experiments using the RAVE method with different groups of respondents, ranging from teacher trainees, psychology students, students of sociology, active teachers, the public at large etc, in Sweden and elsewhere. The experiments have been carried out in other cultural contexts, in the Seychelles, in particular, in order to test the generality of the hypothesis regarding a filtering function. Gendered stereotypes are different in the Seychelles; it has been described as a matriarchal society "where women and girls have many advantages over men and boys" (African Development Bank, 2009) which makes the country a suitable reference point for cross-cultural comparisons.

All trials conducted addressing gender stereotyping have supported our hypothesis that linguistic stereotyping acts as a filter. In trials conducted with teacher trainees in Sweden (n = 61), we could show that respondents who listened to the male guise overestimated stereotypical masculine conversational features such as how often the speaker interrupted, how much floor space 'he' occupied, and how often 'he' contradicted his counterpart. On the other hand, features such as signalling interest and being sympathetic were overestimated by the respondents when listening to the female guise.

Results from the Seychelles have strengthened our hypothesis about linguistic stereotyping. Surveys investigating linguistic features associated with gender showed that respondents' (n=46) linguistic gender stereotyping was quite different from that of Swedish respondents. For example, the results from the Seychelles trials showed that floor space and the number of interruptions made were overestimated by the respondents listening to the **female** guise, quite unlike the Swedish respondents, but still in line with our hypothesis since the stereotypes relating to gender puts women in a position where they can e.g. interrupt.

Trials using psychology students (n=101) have similar results. In experiments where students were asked to rate a case character's ('Kim') personality traits and social behaviour, our findings show that the male version of Kim was deemed more unfriendly and a bit careless compared to the female version of Kim, who was regarded to be more friendly and careful. Again, this shows that respondents overestimate aspects that confirm their stereotypic preconceptions.

REFERENCES

African Development Bank (2009). *Seychelles. Gender Socialisation in the Home: Its impact on boy's achievement in primary and secondary schools.* Human Development Department (OSHD).

Bargh, J.A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71,* 230-244.

Paul Boersma & David Weenink (2016): *Praat: doing phonetics by computer* [Computer program]. Version 6.0.21, retrieved from http://www.praat.org/

Crawford, M. (1995). *Talking Difference: On Gender and Language.* London: Sage. Darcy.

Lambert, W.E., R.C. Hodgson, R.C. Gardner & S. Fillenbaum. (1960). Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology 60*(1), 44-51.

Talbot, M. (2003). Gender stereotypes. Reproduction and challenge. In J. Holmes, & M. Meyerhoff (eds), *The Handbook of Language and Gender*, 468-486. Oxford: Blackwell

Digital Humanities in the Nordic Countries 2018 conference

Final abstract (long paper, 20 min)
Reima Välimäki
Postdoctoral research fellow
Cultural History, University of Turku
Consortium Profiling Premodern Authors (Propreau)

Refutatio errorum – authorship attribution on a late-medieval antiheretical treatise.

Since Peter Biller's attribution of the *Cum dormirent homines* (1395) to Petrus Zwicker, perhaps the most important late medieval inquisitor prosecuting Waldensians, the treatise has become a standard source on the late medieval German Waldensianism. There is, however, another treatise, known as the *Refutatio errorum*, which has gained far less attention. In my dissertation (2016) I proposed that similarities in style, contents, manuscript tradition and composition of the *Refutatio errorum* and the *Cum dormirent homines* are so remarkable that Petrus Zwicker can be confirmed as the author of both texts. The *Refutatio* exists in four different redactions. However, the redaction edited by J. Gretser in the 17th century, and consequently used by modern scholars, does not correspond to the earlier and more popular redaction that is in the majority of preserved manuscripts.

In the paper I will add a new element of verification to Zwicker's authorship: machine-learning-based computational authorship attribution applied in the digital humanities consortium Profiling Premodern Authors (University of Turku, 2016–2019). In its simplest form, the authorship attribution is a binary classification task based on textual features (word uni/bi-grams, character n-grams). In our case, the classifications are "Petrus Zwicker" (based on features from his known treatise) and "not-Zwicker", based on features from a background corpus consisting of medieval Latin polemical treatises, sermons and other theological works. The test cases are the four redactions of the *Refutatio errorum*. Classifiers used include a linear Support Vector Machine and a more complex Convolutional Neural Network. Researchers from the Turku NLP group (Aleksi Vesanto, Filip Ginter, Sampo Pyysalo) are responsible for the computational analysis.

The paper contributes to the conference theme *History.* It aims to bridge the gap between authorship attribution based on qualitative analysis (e.g. contents, manuscript tradition, codicological features, palaeography) and computational stylometry. Computational methods are treated as one tool that contributes to the difficult task of recognising authorship in a medieval text. The study of author profiles of four different redactions of a single work contributes to the discussions on scribes, secretaries and compilers as authors of medieval texts (e.g. Reiter 1996, Minnis 2006, Connolly 2011, Kwakkel 2012, De Gussem 2017).

Bibliography:
Biller, Peter. "The Anti-Waldensian Treatise Cum Dormirent Homines of 1395 and its Author." In *The Waldenses, 1170-1530: Between a Religious Order and a Church*, 237–69. Variorum Collected Studies Series. Aldershot: Ashgate, 2001.

Connolly, Margaret. “Compiling the Book.” In The Production of Books in England 1350-1500, edited by Alexandra Gillespie and Daniel Wakelin, 129–49. Cambridge Studies in Palaeography and Codicology 14. Cambridge ; New York: Cambridge University Press, 2011.

De Gussem, Jeroen. “Bernard of Clairvaux and Nicholas of Montiéramey: Tracing the Secretarial Trail with Computational Stylistics.” *Speculum* 92, no. S1 (2017): S190–225. https://doi.org/10.1086/694188.

Kwakkel, Erik. “Late Medieval Text Collections. A Codicological Typology Based on Single-Author Manuscripts.” In *Author Reader Book: Medieval Authorship in Theory and Practice*, edited by Stephen Partridge and Erik Kwakkel, 56–79. Toronto: University of Toronto Press, 2012.

Reiter, Eric H. “The Reader as Author of the User-Produced Manuscript: Reading and Rewriting Popular Latin Theology in the Late Middle Ages.” *Viator* 27, no. 1 (1996): 151–70. https://doi.org/10.1484/J.VIATOR.2.301125.

Minnis, A. J. “Nolens Auctor Sed Compilator Reputari: The Late-Medieval Discourse of Compilation.” In *La Méthode Critique Au Moyen Âge*, edited by Mireille Chazan and Gilbert Dahan, 47–63. Bibliothèque d’histoire Culturelle Du Moyen âge 3. Turnhout: Brepols, 2006.

Välimäki, Reima. “The Awakener of Sleeping Men. Inquisitor Petrus Zwicker, the Waldenses, and the Retheologisation of Heresy in Late Medieval Germany.” PhD Thesis, University of Turku, 2016.

From crowdsourcing cultural heritage to citizen science: how the Danish National Archives 25-year-old transcription project is meeting digital historians

Bárbara Revuelta-Eugercios[1,2], bre@sa.dk
Nanna Floor Clausen[1], nc@sa.dk
Katrine Tovgaard-Olsen[1], ket@sa.dk

[1] Rigsarkivet (Danish National Archives)
[2] Saxo Institute (University of Copenhagen)

Keywords: crowdsourcing, digital history, citizen science, historical records, archive

Extended abstract

The Danish National Archives have the oldest crowdsourcing project in Denmark, with more than 25 million records transcribed that illuminate the lives and deaths of Danes since the early 18th century. Until now, the main group interested in creating and using these resources has been amateur historians and genealogists. However, it has become clear that the material also holds immense value to historians, armed with the new digital methods. The rise of citizen science projects show, likewise, an alternative way, with clear research purposes, of using the crowdsourcing of cultural heritage material. How to reconcile the traditional crowd-centered approach of the existing projects, to the extent that we can talk about co-creation, with the narrowly-defined research questions and methodological decisions researchers required? How to increase the use of these materials by digital historians without losing the projects' core users?

This article articulates how the Danish National Archives (*Rigsarkivet*) are answering these questions. In the first section, we discuss the tensions and problems of combining crowdsourcing digital heritage and citizen science; in the second, the implications of the crowd-centered nature of the project in the incorporation of research interests; and in the third one, we present some strategies adopted to successfully attract digital historians to work on this material.

Crowdsourcing cultural heritage: for the public and for the humanists

In the last decades, GLAMs (galleries, libraries, archives and museums) have been embarked in digitalization projects to broaden the access, dissemination and appeal of their collections, as well as enriching them in different ways (tagging, transcribing, etc.), as part of their institutional missions. Many of these efforts have included audience or community participation, which can be loosely defined as either crowdsourcing or activities that predate or conform to its standard definition. Howe's (2006) first business-related definition describes it as "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call" (Ridge 2014). However, the key feature that differentiates these crowdsourcing cultural heritage projects is that the work the crowd performs has never been undertaken by employees. Instead, they co-create new ways for the collections to be made available, disseminated, interpreted, enriched and enjoyed that could never had been paid for within their budgets.

These projects often feature "the crowd" at both ends of the process: volunteers contribute to improve access to and availability of the collections, which in turn will benefit the general public from which volunteers are drawn. In the process, access to the digital cultural heritage material is democratized and facilitated, transcribing records, letters, menus, tagging images, digitizing new material, etc. As a knock-on effect, the research community can also benefit, as the new materials open up possibilities for researchers in the digital humanities, which would never have achieved the transcription of millions of records within their financially limited projects.

At the same time, there has been a strand of academic applications of crowdsourcing in Humanities projects (Dunn and Hedges 2014). These initiatives fall within the so-called citizen science projects, which are driven by researchers and narrowly defined to answer a research question, so the tasks performed by the volunteers are lined up to a research purpose. Citizen science or public participation on scientific research, that emerged out of natural sciences projects in the mid-1990s (Bonney et al 2009), has branched out to meet the Humanities, building on a similar utilization of the crowd, i.e. institutional digitalization projects of cultural heritage material. In particular, archival material has been a rich source for such endeavours: weather observations from ship logs in *Old Weather* (Blaser 2014), Benthan's works in *Transcribe Bentham* (Causer & Terras 2014) or restaurant menus on *What's on the menu* (2014). While some of them have been carried out in cooperation with the GLAMs responsible for those collections, the new opportunities opened up for the digital humanities allow these projects to be carried out by researchers independently from the institutions that host the collections, missing a great opportunity to combine interests and avoid duplicating work.

Successfully bringing a given project to contribute to crowdsourcing cultural heritage material and citizen science faces many challenges. First, a collaboration needs to be established across at least two institutional settings – a GLAMs and a research institution- that have very different institutional aims, funding, culture and legal frameworks. GLAMs have foundational missions, which aim at the general public, so the research community is only a tiny percentage of its users. Any institutional research they undertake on the collections is restricted to particular areas or aspects of the collections and institutional interest which, on the other hand, is less dependent on external funding. The world of Academia, on the other hand, has a freer approach to formulating research questions but is often staffed with short-term positions and projects, time-constraints and a need of immediacy of publication and the ever-present demand for proving originality and innovation.

Additionally, when moving from cultural heritage dissemination to research applications, a wide set of issues also come into view: the boundaries between professional and lay expertise, the balance of power in the collaboration between the public, institutions and researchers, ethical concerns in relation to data quality and data property, etc. (Riesh 2014, Shirk et al 2012).

The Danish National Archives crowd-centered 25-year-old project

The Danish National Archives are dealing with the challenge of how to incorporate a more citizen-science oriented approach and attract historians (and digital humanists) to work with the existing digitized sources while maintaining its commitment to the volunteers. This challenge is of a

particular difficulty in this case because not only the interests of the archives and researchers need to align, but also those of the "crowd" itself, as volunteers have played a major role in co-creating crowdsourcing for 25 years.

The original project, now the Danish Demographic Database, DDD, (www.ddd.dda.dk), is the oldest "crowdsourcing project" in the country. It started in 1992 thanks to the interest of the genealogical communities in coordinating the transcription of historical censuses and church books. (Clausen & Jørgensen 2000). From its beginning, the volunteers were actively involved in the decision-making process of what was to be done and how, while the Danish National Archives were in charge of coordination and dissemination functions. Thus, there has been a dual government of the project and a continuous negotiation of priorities, in the form of, a coordination committee, which combines members of the public and genealogical societies as well as DNA's staff.

This tradition of co-creation has shaped the current state of the project and its relationship to research. The subsequent Crowdsourcing portal, CS, (https://cs.sa.dk/), which started in 2014 with an online interface, broadened the sources under transcription and the engagement with volunteers (in photographing, counselling, etc.), and maintains a strong philosophy of serving the volunteers' wishes and interests, rather than imposing particular lines. Crowdsourcing is seen as more than a framework for creating content: it is also a form of engagement with the collections that benefits both audiences and archive. However, it has also introduced some citizen-science projects, in which the transcriptions are intended to be used for research (e.g. the Criminality History project).

Digital history from the crowdsourced material: present and future

While *Arkivalieronline*, the collection of scanned images freely available online, is widely used among amateur historians, genealogists, historians and the public alike, the crowdsourcing projects are only widely used in genealogist and amateur historian circles. Some of the ways in which we are trying to reach professional historians and students are the following:

1. Disseminating the collections to different academic communities. On the one hand, in specialized fields as family history, demography or economic history (European Social History Conference) but also in larger digital humanities. From the beginning DDD, like other projects featuring individual-named tabulated material, participated in the Associate for History and Computing meetings. Lately, these efforts have been renewed by re-joining the field of digital humanities through participation in national events (Dighumlab) and international conferences such as DHN 2018.
2. Providing free extractions for individual users, through our own webpage but also through the North Atlantic Population Project (NAPP.org) at IPUMS, University of Minnesota, which has resulted in more than 50 articles being produced (see http://www.ddd.dda.dk/publikationer.html)
3. Participating in large-scale projects with university partners (particularly, with the University of Copenhagen) to actively pursue research with the collections. As part of the SHiP (Studies of Health in Port Cities) network, researchers at the DNA and the university collaborate in studying 19th century epidemiology patterns. The Link-Lives project is a partnership with the

university and Copenhagen City Archive to link people through the different transcribed sources available for research purposes.

4. Supporting research on the collections by providing training on how to use them. The lack of technical abilities within History faculties or student bodies in Denmark largely explains their under-utilization. The *Rigsarkivet* Digital History Methods Labs is a pilot project targeted at University students to disseminate the collections by providing a basic methodological training on how to use them for historical research. It consists on a series of small workshops where students learn the methods to address a research question using an extraction from our collections.
5. Expanding the focus of some of the traditional genealogist-driven crowdsourcing projects to incorporate a citizen-science approach. For example, the Death certificate project, initiated and run still today by volunteers, to whom DNA mostly facilitates their work (i.e., providing equipment for photographing, uploading and setting up the transcription project). However, the interest that the project has arisen among some historians and epidemiologists makes it advisable to try to bring some research considerations into play. However, this cannot be result in a researcher take-over, being paramount for the survival of the project as well as the preservation of the community to respect the citizens's ownership of the project and invite them to collaborate to make interests align.

The challenges described are not necessarily novel, as GLAM institutions (and in particularly archives) in the Nordic countries have similar collections, are involved in similar crowdsourcing projects and have similar communities that could be attracted. Thus, while some of the actions respond to specific Danish aspects (such as the reduced presence of the fields of digital history and historical demography in Denmark), most of them are already being implemented or could be implemented in other countries. Thus, the existence of partnerships university-archives in many Nordic countries around transcribed historical records (National archives of Norway and Sweden and universities of Tromsø and Umeå, for example) could very well also be used to form international inter-institutional alliances that could bring together efforts, reach a wider community and minimize investments in order to boost/liberate the potential of both the crowd and the collections.

References

Blaser, L., 2014 "*Old Weather*: approaching collections from a different angle" in Ridge (ed) *Crowdsourcing our Cultural Heritage*, Ashgate, 45-56.

Bonney et al. 2009. Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. Center for Advancement of Informal Science Education (CAISE), Washington, DC

Clausen, N.C and Marker, H.J., 2000, "The Danish Data Archive" in Hall, McCall, Thorvaldsen *International historical microdata for population research*, Minnesota Population Center . Minneapolis, Minnesota, 79-92,

Causer, T. and Terras, M. 2014, "'Many hands make light work. Many hands together make merry work': *Transcribe Bentham* and crowdsourcing manuscript collections", in Ridge (ed) *Crowdsourcing our Cultural Heritage*, Ashgate, 57-88.

Dunn, S. and Hedges, M. 2014"How the crowd can surprise us: Humanities crowd-sourcing and the creation of knowledge", in Ridge (ed) *Crowdsourcing our Cultural Heritage*, Ashgate, 231-246.

Howe, J. 2006, "The rise of crowdsourcing", Wired, June.

Ridge, M. 2014, "Crowdsourcing our cultural heritage: Introduction", in Ridge (ed) *Crowdsourcing our Cultural Heritage*, Ashgate, 1-16.

Riesch, H., Potter, C., 2014. Citizen science as seen by scientists: methodological, epistemological and ethical dimensions. Public Understanding of Science 23 (1), 107–120

Shirk, J.L et al, 2012. Public participation in scientific research: a framework for deliberate design. Ecology and Society 17 (2),

Sculpting Time: Temporality in the Language of Finnish Socialism, 1895–1917

At the beginning of the twentieth century, the Grand Duchy of Finland had the largest socialist party with parliamentary representation in Europe (Eley 2002, 66). The breakthrough of Finnish socialism has not yet been analyzed from the perspective of 'temporality'—i.e., the way human beings experience time. This paper examines socialist experiences and expectations by asking three questions: 1) What are the key differences between socialist temporality and non-socialist temporalities in the early twentieth century? 2) What kinds of meanings do socialists attach to the past, present, and future? and 3) Do the actual revolutionary events in Finland—such as the General Strike in the autumn of 1905 and the Russian Revolution in March 1917—change the socialist perception of time in any way?

The data set consists of digitized newspapers published in Finland from 1895 to 1917. The raw text files of all the Finnish newspapers will be downloaded from the National Library of Finland (Pääkkönen et. al. 2016) and lemmatized with the LAS command-line tool (Mäkelä 2016). Four different sub-corpora representing the main political languages of the time will be constructed based on the political affiliation of the given newspaper: socialist, conservative-nationalist, liberal-nationalist, and Christian. The political affiliation will be determined using earlier research on Finnish newspapers (Tommila 1987).

The methodology combines traditional conceptual-historical approaches with the corpus-linguistic methods of keyness, collocation, and key collocation. Conceptual historians have studied the life spans of certain fundamental key concepts through the qualitative analysis of the concepts' temporal layers and linguistic contexts—e.g., the variation of parallel and counter concepts attached to key concepts in different times and places (Koselleck 2004). Corpus linguists, on the other hand, have used computational methods to find patterns in vast text collections (Baker 2006).

The keyness method can be used to show all the words that are used more frequently than expected by pure chance in socialist texts compared to non-socialist texts. Words connected to temporality will be then chosen manually from this keyword list for further analysis. Thus, there is no predetermined operationalization for temporal vocabulary; rather, a bottom-up approach is used. Keyness analysis will shed light on the question of how the socialist formulation of the past, present, and future differed from competing non-socialist perceptions of time.

Collocates are words that appear more frequently than expected by pure chance in close proximity to the search word. The collocates of socialist temporal words found in the keyness analysis will be

quantified in order to find tentative information on the meanings that socialists gave to the past, present, and future in their political language.

Finally, the variation of socialist temporality through time will be analyzed by quantifying the collocates of temporal words *before* and *after* major political upheavals in 1905 and 1917. In practice, all the words appearing in a window of five words to the left or right of the studied temporal word will be collected and combined into one 'post-revolutionary' mini-corpus, which will then be compared with all the 'pre-revolutionary' collocates of the same noun. While collocates can reveal the semantic content of a temporal concept, this method—which could be called key collocation—reveals semantic differences in the use of temporal concepts at different times.

The underlying hypothesis is that identifying the changes in socialist temporality will improve our historical understanding of the political ruptures in Finland in the early twentieth century. The results of the analysis will be uploaded to the GitHub repository (https://github.com/rt80119/dhn2018), and they will be compared to Reinhart Koselleck's famous theory of 'temporalization of concepts' – expectations towards the future supersede experiences of the past in modernity (Koselleck 2004, 9–25, 255–275).

The paper will eventually form a part of my ongoing dissertation project, which merges a close reading of archival sources with computational distant reading of digital materials, thus producing a macro-scale picture of the political language of Finnish socialism.

Bibliography

Baker, Paul 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Eley, Geoff 2002. *Forging Democracy. The History of the Left in Europe, 1850–2000*. Oxford: Oxford University Press.

Koselleck, Reinhart 2004. *Futures Past. On the Semantics of Historical Time*. New York: Columbia University Press.

Mäkelä, Eetu 2016. "LAS: An Integrated Language Analysis Tool for Multiple Languages". *Journal of Open Source Software* 1(6), 35, DOI: 10.21105/joss.00035.

Pääkkönen, Tuula, Kervinen, Jukka, Nivala, Asko, Kettunen, Kimmo & Mäkelä, Eetu 2016. "Exporting Finnish Digitized Historical Newspaper Contents for Offline Use". *D-Lib Maganize* vol. 22, no 7/8, DOI: 10.1045/july2016-paakkonen.

Tommila, Päiviö (ed.) 1987. *Suomen lehdistön historia 2. Sanomalehdistö suurlakosta talvisotaan*. Helsinki: Kustannuskiila.

Broken data and unexpected research questions

Minna Ruckenstein, University of Helsinki

Recent research introduces the concept-metaphor of "broken data", suggesting that digital data might be broken and fail to perform, or be in need of repair (Pink et al 2018). Concept-metaphors, anthropologist Henrietta Moore (1999, 16; see also Moore 2004) argues, are domain terms that "open up spaces in which their meanings – in daily practice, in local discourses and in academic theorizing – can be interrogated." By doing so, concept-metaphors become defined contextually in practice; they are not meant to be foundational concepts, but work as partial and perspectival framing devices.

In this presentation, the concept-metaphor of broken data is discussed in relation to the open data initiative, *Citizen Mindscapes*, an interdisciplinary project that contextualizes and explores a Finnish-language social media data set ('*Suomi24*', or Finland24 in English), consisting of tens of millions of messages and covering social media over a time span of 15 years (see, Lagus et al 2016). The aim of taking advantage of a concept-metaphor in a data-related study is to arrange and provoke ideas and open a conceptual domain within which facts, connections and relationships are identified and imagined. The role of the broken data metaphor in this discussion is to examine the implications of breakages and consequent repair work in data-driven initiatives that take advantage of secondary data. Moreover, the concept-metaphor can sensitize us to consider the less secure and ambivalent aspects of data work. By focusing on how data might be broken, we can highlight misalignments between people, devices and data infrastructures, or bring to the fore the failures to align data sources or data uses with the everyday.

As Pink et al (2018) suggest the metaphorical understanding of digital data, aiming to underline aspects of data brokenness, brings together various strands of scholarly work, highlighting important continuities with earlier research. Studies of material culture explore practices of breakage and repair in relation to the materiality of objects, for instance by focusing on art restoration (Dominguez Rubio 2016), or car repair (Dant 2010). Drawing attention to the fragility of objects and temporal decay, these studies underline that objects break and have to be mended and restored. When these insights are brought into the field of data studies, the materiality of databases,

platforms and software become a concern (Tanweer et al 2016), emphasizing aspects of brokenness and following repair work in relation to digital data (Pink et al 2018).

In the science and technology studies (STS), on the other hand, the focus on 'breakages' has been studied in relation to infrastructures, demonstrating that it is through instances of breakdown that structures and objects, which have become invisible to us in the everyday, gain a new kind of visibility. The STS scholar Stephen Jackson expands the notion of brokenness to more everyday situations and asks 'what happens when we take erosion, breakdown, and decay, rather than novelty, growth, and progress, as our starting points in thinking through the nature, use, and effects of information technology and new media?' (2014: 174). Instances of data breakages can be seen in light of mundane data arrangements, as a recurring feature of data work rather than an exceptional event (Pink et al 2018; Tanweer et al 2016).

In order to concretize further the usefulness of the concept-metaphor of broken data, I will describe how identifying instances of breakage in the data set and repair in the data work can generate new and unanticipated research questions. In particular, I will highlight the role of spam bots, computer programs specifically designed for generating spam messages, in digital work. In the collaborative Citizen Mindscapes initiative, discussing the gaps, or possible anomalies in the data led to conversations concerning the production of data, deepening our understanding of the human and material factors at play in processes of data generation. As described below, the waste of the data world, spam messages, could also be seen as a resource in terms of everyday digital innovation (Tanweer et al 2016).

Working with spam

The Suomi24 data was generated by the media company, Aller; the data silently resided on the servers until the company decided to open the proprietary data for research purposes (see Lagus et al 2016). In the past two years, the *Citizen Mindscapes* -initiative, particularly researchers experienced in working with large data sets, have been cleaning the data in order to make it ready for computational work. The aim is to build a methodological toolbox that researchers, who do not possess computational skills, but are interested in using digital methods in the social scientific inquiry, can benefit from. This entails, for instance, developing user interfaces that narrow down the huge data set and allow to access the data with topic-led perspectives.

The ongoing work has alerted the research collective to breakages of data, raising more general questions about the origins and nature of data (Pink et al 2018). The research report that details and contextualizes the Suomi24 data pays attention to the writers of the social media community as producers of the data; the moderation practices of the company are described to demonstrate how they shape the data set by filtering words and terms, or certain kinds of messages, for instance, advertisement or messages containing sensitive personal information (Lagus et al 2016). When the data work identifies gaps, errors and anomalies in the data, it reveals that data might be broken and discontinuous due to human or technological forces: infrastructure failures, trolling, or automated spam bots.

We have repeatedly used the visual information of gaps in the data (see Figure 1) as a conversation opener with the social media company's employees. We learned that the 2004-2005 is probably a technical error in the database retrieval. The anomaly in the data volume in July 2009 was first defined as a spam bot by the employees (Pink at al, 2018). Later, however, one of the moderators of the company suspected that it could not have been a spam bot after all. The data set was not supposed to contain spam in such quantities, because the data was already cleaned by the programmers.

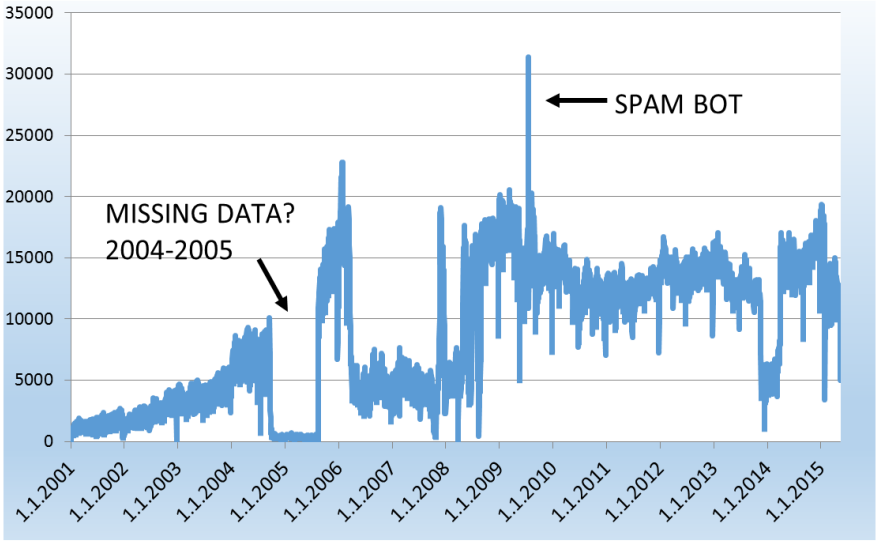


Figure 1: Identified gaps and breakages in the Suomi24-data

In January 2018, we started exploring the July 2009 peak in a more consistent manner. A whole new area of research questions started to emerge about automation and spam bots. Importantly, the

spam bots have an online agency of their own; bots are searching for wikis, blogs and forums that they can use to submit spam. In the Suomi24 forum, the spam messages involve techniques of targeted advertisement, and at times it might be hard to tell human-generated posts apart from automated ones. Other spam messages are not meant to be read by humans at all, but they are posted to increase the number of links to a particular website in order to boost its search engine ranking.

One of the programmers of the company describes the "cat and mouse chase" with the spammers and spam bots; in the past ten years spam has been for him one of the biggest problems of the discussion forum. He feels that only recently they have finally started to master the spam by using machine-enabled filtering. One of the latest incidents of automated posting of links was to boost the search engine ranking of a sport-related event. Occasionally the link posting is also done by humans. Based on IP addresses, the human-generated spamming is mainly produced from India, frequently referred to in the press as "the spam capital of the world".

From the perspective of the broken data metaphor, spam bots raise further questions about brokenness and repair work by paying attention to how the discussion forum, and the data that it generates, is kept clean by filtering it manually and automatically. We now know for sure that the peak in the data on the 17th of July 2009 was a spam bot: the amount of messages on that day was 32 033. Of these messages 17 850 contained the following text: "This message has been removed by admin." From the perspective of data work, analyzing the removal messages, in the context of the discussion forum, might, for instance, tell us about the temporal rhythms of spammers or which conversation threads are more likely to be infected with unwanted content. From the programmer's perspective, on the other hand, the cleaning work is active development of new filters and tools that search the discussion forum in order to identify harmful or rubbish content. For a programmer this work can be quite exciting and enjoyable, trying to get on top of the spammers and being one step ahead. Spam bots call for improvisation and creativity, highlighting the role of repair work as an important resource for knowledge production and innovation (Tanweer et al 2016).

Concluding remarks

The broken data concept metaphor calls for paying more attention to the incomplete, fractured and changing character of digital data. The particular example used to highlight the shifting character of data focused on a peak in a data visualization identified as a spam bot. Acknowledging the incomplete nature of digital data in itself is of course nothing new, researchers are well aware of

their data lacking perfection. With growing uses of secondary data, however, the ways in which data is broken and incomplete might not be known beforehand, underlining the need to explore brokenness and the consequent work of repair. In the case of Suomi24, the data breakages suggest that we need to actively question data production and the diverse ways in which data are adapted for different ends by practitioners. As our *Citizen Mindscapes* collaboration suggests, the production of data is permeated by moments of breakdown and repair that call for a richer understanding of everyday data work and data practices. The intent of this paper has been to suggest that a focus on data breakages is an opportunity to learn about digital work, and to account for how data breakages and related uncertainties challenge linear and too confident stories about data work.

References

Bell, G. (2015). 'The secret life of big data'. In Data, now bigger and better! Eds. T. Boellstorf and B. Maurer. Publisher: Prickly Paradigm Press,7-26

Dant, T., 2010. The work of repair: Gesture, emotion and sensual knowledge. *Sociological Research Online*, 15(3), p.7.

Domínguez Rubio, F. (2016) 'On the discrepancy between objects and things: An ecological approach' *Journal of Material Culture* 21(1): 59–86

Jackson, S.J. (2014) 'Rethinking repair' in T. Gillespie, P. Boczkowski, and K. Foot, eds. *Media Technologies: Essays on Communication, Materiality and Society*. MIT Press: Cambridge MA

Lagus, K. Pantzar, M. Ruckenstein, M. and Ylisiurua M. (2016) Suomi24: Muodonantoa aineistolle. The Consumer Society Research Centre. Helsinki: Faculty of Social Sciences, University of Helsinki.

Moore, H (1999) Anthropological theory at the turn of the century in H. Moore (ed) *Anthropological Theory Today*. Cambridge: Polity Press, pp. 1-23.

Moore, H (2004) Global anxieties: concept-metaphors and pre-theoretical commitments in anthropology. *Anthropological theory*, 4(1), 71-88.

Pink, S., Ruckenstein, M., Willim, R., & Duque, M. (2018). Broken data: Conceptualising data in an emerging world. *Big Data & Society*, *5*(1), 2053951717753228.

Tanweer, A., Fiore-Gartland, B., and Aragon, C. (2016). Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process. *Information, Communication & Society*, 19(6), 736-752.

A long way? Introducing digitized historical newspapers in school, a case study from Finland

Author

Inés Matres García del Pino
Faculty of Arts / Department of Ethnology
University of Helsinki
Ines.Matres@helsinki.fi
orcid.org/0000-0002-4544-4946

Abstract

During 2016/17 two Finnish newspapers, from their first issue to their last, were made available to schools in eastern Finland through the digital collections of the National Library of Finland (http://digi.kansalliskirjasto.fi). This paper presents the case study of one upper-secondary class making use of these materials. Before having access to these newspapers, the teachers in the school in question had little awareness of what this digital library contained. The initial research questions of this paper are whether digitised historical newspapers can be used by school communities, and what practices they enable. Subsequently, the paper explores how these practices relate to teachers' habits and to the wider concept of literacy, that is, the knowledge and skills students can acquire using these materials. To examine the significance of historical newspapers in the context of their use today, I rely on the concept of 'practice' defined by cultural theorist Andreas Reckwitz as the "use of things that 'mould' activities, understandings and knowledge".

To correctly assess practice, I approached this research through ethnographic methods, constructing the inquiry with participants in the research: teachers, students and the people involved in facilitating the materials. During 2016, I conducted eight in-depth interviews with teachers about their habits, organized a focus group with further 15 teachers to brainstorm activities using historical newspapers, and observed a class of 17-18-year-old students whose literature teacher decided to implement the materials right away. Observing the students' work, hearing their presentations, motivations, and opinions about the materials showed how students explored the historical background of their existing personal, school-related and even professional interests. In addition to the students' projects, I also collected their newspaper clippings and logs of their searches in the digital library. These digital research assets revealed how the digital library that contains the historical newspapers influenced the students' *freedom* to choose a topic to investigate and their capacity to *go deep* in their research.

The findings of this case study build upon, and extend, previous research about how digitized historical sources contribute in upper-secondary education. The way students used historical newspapers and accounts of teachers in interviews revealed similarities with activities using present-day newspapers, already a popular material in Finnish schools. Additionally, both the historicity and the form of presentation of newspapers in a digital library confer unique attributes upon these materials: they allow students to explore the historical background of their research interests, discover change across time, verbalize their research ideas in a concrete manner, and train their skills in distant and close reading to manage large amounts of digital content. In addition to these positive attributes that connect with learning goals set by teachers, students also tested the limits of these materials. The lack of metadata in articles or images, the absence of colour in materials that originally have it, or the need for students to be mindful

of how language has changed since the publication of the newspapers are constrains that distinguish digital libraries from resources, such as web browsers and news sites, that are more familiar to students. Being aware of these positive and negative affordances, common to digital libraries containing historical newspapers and other historical sources, can support teachers in providing their students effective guidelines when using this kind of materials.

This use case demonstrates that digitized historical sources in education can do more than enable students to "follow the steps of contemporary historians", as research has previously established. In addition to fitting in history curriculum, these newspapers occupy a place between history and media education. The objective of media education in school –regardless of the technological underpinnings of a single medium, which change rapidly in this digital age– aims at enabling students to reflect on the processes of media production and consumption. The contribution of digitized historical newspapers to this subject is acquainting students with processes of media preservation and heritage. However, it could still be a long way until teachers adopt these aspects in their plans. It is necessary to acknowledge the trajectory and agents involved in the work of introducing newspapers in education, since the 1960s. This task not only consisted of facilitating access to newspapers, but also developing teaching plans and a common understanding of media education in schools.

In addition to uncovering an aspect of digital cultural heritage that is relevant for the school community today, this paper raises awareness among the cultural heritage community, especially national libraries, about the diversity in the uses and users of their collections, especially in a time when the large-scale digitization of special collections is generalizing access to materials traditionally considered for academic research.

Keywords: newspapers in education, historical newspapers, digital libraries, upper-secondary school, case study, national libraries.

Selected bibliography:

Buckingham, D. (2003). Media education: literacy, learning, and contemporary culture. Polity Press.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). Digital_Humanities. MIT Press.

Gooding, P. (2016). Historical newspapers in the Digital Age: 'Search All About It!' Routledge.

Lévesque, S. (2006). Discovering the Past: Engaging Canadian Students in Digital History. Canadian Social Studies, 40(1).

Martens, H. (2010). Evaluating Media Literacy Education: Concepts, Theories and Future Directions. Journal of Media Literacy Education, 2(1).

Nygren, T. (2015). Students Writing History Using Traditional and Digital Archives. Human IT, 12(3), 78–116.

Reckwitz, A. (2002). Toward a Theory of Social Practices: A Development in Culturalist Theorizing. European Journal of Social Theory, 5(2), 243–263.

Towards an Approach to Building Mobile Digital Experiences For University Campus Heritage & Archaeology

Ethan Watrall
Assistant Professor, Department of Anthropology
Associate Director, MATRIX: The Center for Digital Humanities & Social Sciences
Director, Cultural Heritage Informatics Initiative

The spaces we inhabit and interact with on a daily basis are made up of layers of cultural activity that are, quite literally, built up over time. While museum exhibits, archaeological narratives, and public programs communicate this heritage, they often don't allow for the public to experience interactive, place-based, and individually driven exploration of content and spaces. Further, designers of public heritage and archaeology programs rarely explore the binary nature of both the presented content and the scholarly process by which the understanding of that content was reached. In short, the scholarly narrative of material culture, heritage, and archaeology is often hidden from public exploration, engagement, and understanding. Additionally, many traditional public heritage and archaeology programs often find it challenging to negotiate the balance between the voice and goals of the institution and those of communities and groups. In recent years, the maturation of mobile and augmented reality technology has provided heritage institutions, sites of memory and memorialization, cultural landscapes, and archaeological projects with interesting new avenues to present research and engage the public. We are also beginning to see exemplar projects that suggest fruitful models for moving the domain of mobile heritage forward considerably.

University campuses provide a particularly interesting venue for leveraging mobile technology in the pursuit of engaging, place-based heritage and archaeology experiences. University campuses are usually already well traveled public spaces, and therefore don't elicit the same level of concern that you might find in other contexts for publicly providing the location of archaeological and heritage sites and resources. They have a built in audience of alumni and students eager to better understand the history and heritage of their home campus. Finally, many university campuses are starting to seriously think of themselves as places of heritage and memory, and are developing strategies for researching, preserving, and presenting their own cultural heritage and archaeology.

It is within this context that this paper will explore a deeply collaborative effort at Michigan State University that leverages mobile technology to build an interactive and place-based interpretive layer for campus heritage and archaeology. Driven by the work of the Michigan State University Campus Archaeology Program, an internationally recognized initiative that is unique in its approach to campus heritage, these efforts have unfolded across a number of years and evolved to meet the ever changing need to present the rich and well studied heritage and archaeology of Michigan State University's historic campus.

Ultimately, the goal of this paper is not only to present and discuss the efforts at Michigan State University, but to provide a potential model for other university campuses interested in leveraging mobile technology to produce engaging digital heritage and archaeology experiences.

Art of the Digital Natives and Predecessors of Post-Internet Art

Raivo Kelomees
Senior researcher
Estonian Academy of Arts
E-mail: offline@online.ee

Keywords: digital natives, post-internet art, new aesthetics, history of digital art

Abstract

The aim of the presentation is to pay an homage to the first generation of internet-based artists and to show, how their creative contribution is partially lost and how it became part of the art history. Also, it is of interest to follow the debate around post-internet art, which representatives claim their relation to the net.art. More specifically, I would like to point to and analyse exhibition "ARS17: Hello World!" in Museum of Contemporary Art Kiasma, Helsinki and to analyse art of the "digital natives" and how they differ from artists who invaded the digital art arena in the 1990ties.

Introduction

The aim of the presentation is to pay an homage to the first generation of internet-based artists and to show, how their creative contribution is partially lost and how it became part of the art history. Also, it is of interest to follow the debate around post-internet art, which representatives claim their relation to the net.art. More specifically, I would like to point to and analyse exhibition "ARS17: Hello World!" in Museum of Contemporary Art Kiasma, Helsinki (31.03.2017 - 14.01.2018). Also I would like to analyse art of the "digital natives" and how they differ from artists who invaded the digital art arena in the 1990ties.

Now and then: internet art vs. post-internet art

Now that the internet has drowned in social networks and commercial channels, it makes sense to look back at the experiences of net art in the 1990s. This was an era of innocence, eagerness and heroes of a kind, when networks as art were brand new. Net art was both ironic and self-critical.

The essential difference is that the internet environment back then was something special and new, and now it is banal and everyday. In the 1990s, culture and art had to be, figuratively speaking, brought to the internet, settled there and only then was it possible to see how the environment influenced the content, whereas in the current post-digital and post-internet era, the internet environment is like nature: it surrounds us. It has become a channel through which the world reaches us, but it has also become an environment where people live their everyday lives, communicate and express themselves. It has turned into a dominating environment.

The new normal or the digital environment surrounding us has in recent years surprised us, at least in the fine arts, with the internet's content returning to its physical space. Is this due to pressure from the galleries or something else; in any case, it is clearer than ever that the audience is not separable from the habitual space; there is a huge and primal demand for physical or material art.

Christiane Paul (2017, pp. 37-41) in her article "Digital Art Now: The Evolution of the Post-Digital Age" in "ARS17: Hello World!" exhibition catalogue, is critical of the exhibition. Her main message is that all this has been done before. In itself the statement lacks originality, but in the context of the postinternet apologists declaring the birth of a new mentality, the arrival of a new "after experiencing the internet" and "post-digital" generation, it becomes clear that indeed it is rather like shooting fish in a barrel, because art that is critical of the digital and interactive has existed since the 1990s, as have works concerned with the physicalisation of the digital experience.

The background to the exhibition is the discussion over "digitally created" art and the generation related to it. The notion of "digital natives" is related to the post-digital and post-internet generation and the notion of "post-contemporary" (i.e. art is not concerned with the contemporary but with the universal human condition). Apparently for the digital natives, the internet is not a way out of the world anymore, but an original experience in which the majority of their time is spent. At the same time, however, the internet is a natural information environment for people of all ages whose work involves data collection and intellectual work. Communication, thinking, information gathering and creation – all of these realms are related to the digital environment. These new digital nomads travel from place to place and work in a "post-studio" environment.

While digital or new media was created, stored and shared via digital means, post-digital art addresses the digital without being stored using these same means. In other words, this kind of art exists more in the physical space.

"New aesthetics"

Considerable reference also exists in relation to James Bridle's (2011) new aesthetics concept. In short, this refers to the convergence and conjoinment of the virtual and physical world. It manifests itself clearly even in the "pixelated" design of consumer goods or in the oeuvre of sculptors and painters, whose work has emerged from something digital. For example, the art objects by Shawn Smith and Douglas Coupland are made using pixel-blocks (the sculpture by the latter is indeed reminiscent of a low resolution digital image). Analogous works induce confusion, not to say a surprising experience, in the minds of the audience, for they bring the virtual quality of the computerised environment into physical surroundings. This makes the artworks appear odd and surreal, like some sort of mistake, errors, images and objects out of place.

The so-called postinternet generation artists are certainly not the only ones making this kind of art. As an example of this, there is a reference to the abstract stained glass collage of 11,500 pixels by Gerhard Richter (2007) in the Cologne Cathedral. It is supposed to be a reference to his 1974 painting "4096 Farben" (4096 colours), which indeed is quite similar. It is said that Richter did not accept a fee; however, the material costs were covered by donations. And yet the cardinal did not come to the opening of the glasswork, preferring depictions of Christian martyrs over abstract windows, which instead reminded him of mosques (Welt.de 2007).

One could name other such examples inspired by the digital world or schisms of the digital and physical world: Helmut Smits' "Dead Pixel in Google Earth" (2008); Aram Barholl's "Map" (2006); the projects by Eva and Franco Mattes, especially the printouts of Second Life avatars from 2006; Achim Mohné's and Uta Koppi's project "Remotewords" (2007–2011), computer-based instructions printed on rooftops to be seen from Google Maps or satellites or planes (Mohné and Koppi 2007-11). There are countless examples where it is hard to discern whether the artist is deliberately and critically minded towards digital art or rather a representative of the post-digital generation who is not aware and wishes not to be part of the history of digital art.

From the point of view of researchers of digital culture, the so-called media-archaeological direction could be added to this as an inspirational source for artists today. Media archaeology or the examination of previous art and cultural experience signifies, in relation to contemporary media machines and practices, the exploration of previous non-digital cultural devices, equipment, means of communication, and so on, that could be regarded as the pre-history of today's digital culture and digital devices. With this point of view, the "media-archaeological" artworks of Toshio Iwai or Bernie Lubell coalesce. They have taken an earlier "media machine" or a scientific or technical device and created a modern creation on the basis of it.

Then there was the "Ars Electronica" festival (2006) that focused on the umbrella topic "Simplicity", which in a way turned its back on the "complexity" of digital art and returned to the physical space.

Therefore, in the context of digital media based art trends, the last couple of decades have seen many expressions – works, events and exhibitions – of "turning away" from the digital environment that would outwardly qualify as post-digital and postinternet art.

The most significant mistake that is made regarding new media-based art is to see it as a medium in the sense of a mediator, that it conveys some kind of other reality, translated through a digital code, information of analogous reality, and is then re-mediated. A certain context arises from the characteristics of the phenomenon itself: such as its materialism and technical qualities. The digital environment is not just the transfer medium, the mediator medium, the re-mediator. Various art forms emerging on these platform (interactive art, net art, software art, telecommunicative art, bioart

and other hybrid formats) also possess their own set of rules, which understandably rely on the character of the digital environment and technology, but are essentially innovative. This is not merely technology as a tool, a medium and a means with which to differently package existing reality; it has created a different kind of playing field where the previous conventions of physical art and reality are no longer valid. At the same time, this field requires some technical knowledge.

The 1990s could be characterised as an era of establishing new media centres, with media labs of all kinds, and university new media subunits. However it seems the enthusiasm has waned for artists to acquire the necessary technical knowledge and skills for digital work. This may be because the attraction, edginess and 'sexiness' of new media have decreased, partly because digital technology is everywhere, and partly because purely technical education does not really suit art academies: the cognitive abilities of creative people are limited and they require more intuitive creative practices than technical training involving discrete intellectual abilities. This all constitutes fertile ground for the decisive backlash known as 'post-internet'.

References

1. Paul, C. (2017). Digital Art Now: The Evolution of the Post-Digital Age, in *ARS17: Hello World! Art After Internet*, catalogue, A Museum of Contemporary Art Publication, Kiasma, pp. 37-41.
2. Bridle, J. (2011). *The New Aesthetic: Waving at the Machines, 5. dets 2011.* Retrieved from http://booktwo.org/notebook/waving-at-machines/.
3. Smith, S. Retrieved from http://shawnsmithart.com/images.htm.
4. Coupland, D. Retrieved from http://www.yaean.com/en/blog/2010/07/28/douglas-coupland-orca-sculpture/.
5. Richter, G. (2007). Cologne Cathedral Window. Retrieved from https://www.gerhard-richter.com/en/art/other/glass-and-mirrors-105/cologne-cathedral-window-14890/?p=1.
6. Richter, G. (1974). 4096 Colours. Retrieved from https://www.gerhard-richter.com/en/art/paintings/abstracts/colour-charts-12/4096-colours-6089.
7. Welt.de (2007). Gerhard Richter weist Meisners Kritik zurück. Retrieved from https://www.welt.de/politik/article1148224/Gerhard-Richter-weist-Meisners-Kritik-zurueck.html.
8. Smits, H. (2008). Dead pixel in Google Earth. Retrieved from http://rhizome.org/editorial/2009/mar/30/dead-pixel-in-google-earth-2008-helmut-smits/.
9. Barholl, A. (2006) "Map". Retrieved from http://www.datenform.de/mapeng.html.
10. Mohné, A. and Koppi, U. (2007–2011). "Remotewords". Retrieved from http://www.achimmohne.de/content/remotewords33.html.

Biography

Raivo Kelomees, *PhD (art history)*, artist, critic and new media researcher. Studied psychology, art history, and design in Tartu University and the Academy of Arts in Tallinn. Has published in main cultural and art magazines and newspapers of Estonia since 1985. Book author, “Surrealism” (Kunst Publishers, 1993) and an article collections “Screen as a Membrane” (Tartu Art College proceedings, 2007), "Social Games in Art Space" (Estonian Academy of Arts, 2013). Doctoral thesis „Postmateriality in Art. Indeterministic Art Practices and Non-Material Art“ (Dissertationes Academiae Artium Estoniae 3, 2009).

Abstract, long paper:

A Computational Assessment of Norwegian Literary "National Romanticism"

Ellen Rees, University of Oslo

In this paper, I present findings derived from a computational analysis of texts designated as "National Romantic" in Norwegian literary historiography. The term "National Romantic," which typically designates literary works from approximately 1840 to 1860 that are associated with national identity formation, first appeared decades later, in Henrik Jæger's *Illustreret norsk litteraturhistorie* from 1896. Gudleiv Bø has written extensively about numerous examples of national romanticism in Norwegian literature without probing the term itself to any great extent (Bø 1995, 1998 2006, 2008, 2011). Cultural historian Nina Witoszek has on a number of occasions written critically about the term, claiming that it is misleading because the works it denotes have little to do with larger international trends in Romanticism (see especially Witoszek 2011). Yet, with the exception of a 1985 study by Asbjørn Aarseth, it has never been interrogated systematically within the Norwegian context in the way that other period designations such as "Realism" or "Modernism" have.[1] Nor does Aarseth's investigation attempt to delimit a definitive National Romantic corpus or account for the remarkable disparity among the works that are typically associated with the term. "National Romanticism" is like pornography—we know it when we see it, but it is surprisingly difficult to delineate in a scientifically rigorous way.

Together with members of the project team, I have prepared a corpus of texts that are mentioned in connection with "National Romanticism" in the major histories of Norwegian literature in Norwegian literature. I will discuss briefly some of the logistical challenges associated with preparing this corpus.

This corpus forms the point of departure for a computational analysis employing various text-mining methods in order to determine to what degree the texts most commonly associated with "National Romanticism" share significant characteristics. In the popular imagination, the period is associated with folkloristic elements such as supernatural creatures (trolls, hulders), farming practices (shielings, herding), and folklife (music, rituals) as well as nature motifs (birch trees, mountains). We therefore employ topic modeling in order to map the frequency and distribution of such motifs across time and genre within the corpus. We anticipate that topic modeling will also reveal unexpected results beyond the motifs most often associated with National Romanticism. This process should prepare us to take the next step and, inspired by Matthew Wilkens' recent work generating "clusters" of varieties within twentieth-century U.S. fiction, create visualizations of similarities and differences among the texts in the National Romanticism corpus (Wilkens 2016).

Based on these initial computational methods, we hope to be able to answer some of the following literary historical questions:

---

[1] National romanticism in a central and eastern European context is treated in *National Romanticism: The Formation of National Movements* (Trencsenyi and Kopecek 2007).

- Are there identifiable textual elements shared by the texts in the National Romantic canon?
- What actually defines a National Romantic text as National Romantic?
- Do these texts cluster in a meaningful way chronologically?
- Is "National Romanticism" in fact meaningful as a period designation, or alternately as a stylistic designation?
- Are there other texts that share these textual elements that are not in the canon?
- If so, why? Do gender, class or ethnicity have anything to do with it?

To answer the last two questions, we need to use the "National Romanticism" corpus as a sub-corpus and "trawl-line" within the full corpus of nineteenth-century Norwegian textual culture, carrying out sub-corpus topic modeling (STM) in order to determine where similarities with texts from outside the period 1840–1860 arise (Tangherlini and Leonard 2013). For the sake of expediency, we use the National Library of Norway's Digital Bookshelf as our full corpus, though we are aware that there are significant subsets of Norwegian textual culture that are not yet included in this corpus. Despite certain limitations, the Digital Bookshelf is one of the most complete digital collections of a national textual culture currently available.

For the purposes of DHN 2018, this project might best be categorized as an exploration of cultural heritage, understood in two ways. On the one hand, the project is entirely based on the National Library of Norway's Digital Bookshelf platform, which, as an attempt to archive as much as possible of Norwegian textual culture in a digital and publicly accessible archive, is in itself a vehicle for preserving cultural heritage. On the other hand, the concept of "National Romanticism" is arguably the most widespread, but least critically examined means of linking cultural heritage in Norway to a specifically nationalist agenda.

Preliminary findings indicate that texts associated with "national romanticism" in the literary histories do in fact differ from the larger reference corpus, but not in the ways we expected. We anticipated clusters aligning with familiar national romantic themes (the supernatural, farming practices, folk life and nature motifs), and while these were indeed present, they were not the only markers for national romanticism. A quite different cluster emerged as equally, if not more, dominant, namely that of infatuation. It consists of words like: kiss, cheek, boy, girl, smile, beloved.

These preliminary findings are based on comparisons of word frequency in a target corpus consisting of 78 texts identified as national romantic and a randomly generated reference corpus of 500 books published in Norway between 1830 and 1890. The target corpus is much smaller than our original list because a number of works are shorter texts contained within the same book, such as folk tales or poems. To make the comparison, we generated wordlists for each of the two corpora and a third combined corpus. We then normalized word frequency so that word occurrence is relative to the book it appears in. The target and reference corpora were then aggregated so that they function as lists of word frequency. All words in each book were connected to an average relative frequency, so that the higher the number generated, the more specific the word is to that particular corpus. Each aggregated corpus was divided by the combined corpus, which allows us to see the greatest differences

in word frequency. We reckon that a word with a value of approximately three or higher is specific to a given corpus. Some of these highly frequent words are related to themes, while others are most likely stylistic and related to the genre of the text (literary versus non-fiction, for example).

The methods behind our preliminary findings have raised a number of issues to be worked out and further questions to be asked. For example, both the target corpus and the reference corpus still need to be more rigorously defined. The target corpus was generated from a list of the URNs for texts on the list gleaned from a manual review of the literary histories. This list is problematic because the literary histories often make only tentative associations between—or outright problematize the relationship between—a given text or author and the term "national romanticism." We hypothesize that textbooks used in Norwegian instruction at the secondary school level would be a more appropriate source because they tend to more explicitly link a given text to a given period or style, and we plan to construct a new corpus derived from a manual review of these textbooks. Another reason for using textbooks is that they are much more widely read than literary history books, and thus more central to the formation of the general understanding of the period.

An equally pressing problem with the reference corpus is that it is randomly generated from all books published in the period 1830–1890, rather than being randomly generated from a list of specifically literary books published in the period 1830–1890. We surmise that the inclusion of genres such as, for example, scientific studies, dictionaries, or instruction manuals undermine the validity of our findings, and we thus plan to revise the reference corpus so that it contains only literary texts defined according to standard literary genres.

References:

Jæger, Henrik. 1896. *Illustreret norsk litteraturhistorie.* Bind II. Kristiania: Hjalmar Biglers forlag.

Tangherlini, Timothy R. and Peter Leonard. 2013. "Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research." *Poetics* 41.6: 725–749.

Wilkens, Matthew. 2016. "Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction." *CA: Journal of Cultural Analytics* (online open-access)

Witoszek, Nina. 2011. *The Origins of the "Regime of Goodness": Remapping the Cultural History of Norway.* Oslo: Universitetsforlaget.

Aarseth, Asbjørn. 1985. *Romantikken som konstruksjon: tradisjonskritiske studier i nordisk litteraturhistorie.* Bergen: Universitetsforlaget.

Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914
Poster at DHN 2018 conference
Abstract

Presenters: Hannu Salmi, Asko Nivala, Mila Oiva, Otto Latva – Cultural History, University of Turku

Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914 (OcEx) is a Digging into Data – Transatlantic Platform funded international and interdisciplinary project with a focus on studying spreading of news globally in the nineteenth century newspapers. The project combines digitized newspapers from Europe, US, Mexico, Australia, New Zealand, and the British and Dutch colonies of that time all over the world.

The project examines patterns of information flow, spread of text reuse, and global conceptual changes across national, cultural and linguistic boundaries in the nineteenth century newspapers. The project links the different newspaper corpora, scattered into different national libraries and collections using various kinds of metadata and printed in several languages, into one whole.

~~The project proposes to present a poster in the Nordic Digital Humanities Conference 2018. The project started in June 2017, and the aim of the poster is to present the current status of the project.~~

The research group members come from Finland, the US, the Netherlands, Germany, Mexico, and UK. OcEx's participating institutions are Loughborough University, Northeastern University, North Carolina State University, Universität Stuttgart, Universidad Nacional Autónoma de México, University College London, University of Nebraska-Lincoln, University of Turku, and Utrecht University. The project's 90 million newspaper pages come from Australia's Trove Newspapers, the British Newspapers Archive, Chronicling America (US), Europeana Newspapers, Hemeroteca Nacional Digital de México, National Library of Finland, National Library of the Netherlands (KB), the National Library of Wales, New Zealand's PapersPast, and a strategic collaboration with Cengage Publishing, one of the leading commercial custodians of digitized newspapers.

Objectives
Our team will hone computational tools, some developed in prior research by project partners and novel ones, into a suite of openly available tools, data, and analyses that trace a broad range of language-related phenomena (including text reuse, translational shifts, and discursive changes). Analysing such parameters enables us to characterize "reception cultures," "dissemination cultures," and "reference cultures" in terms of asymmetrical flow patterns, or to analyse the relationships between reporting targeted at immigrant communities and their surrounding host countries.

OcEx will leverage existing relationships and agreements between its teams and data providers to connect disparate digital newspaper collections, opening new questions about historical globalism and modeling consortial approaches to transnational newspaper research. OcEx will take up challenging questions of historical information flow, including:

1. Which stories spread between nations and how quickly?
2. Which texts were translated and resonated across languages?
3. How did textual copying (reprinting) operate internationally compared to conceptual copying (idea spread)?
4. How did the migration of texts facilitate the circulation of knowledge, ideas, and concepts, and how were these ideas transformed as they moved from one Atlantic context to another?
5. How did geopolitical realities (e.g. economic integration, technology, migration, geopolitical power) influence the directionality of these transnational exchanges?
6. How does reporting in immigrant and ethnic communities differ from reporting in surrounding host countries?
7. Does the national organization of digitized newspaper archives artificially foreclose globally-oriented research questions and outcomes?

Methodology

OcEx will develop a semantic interoperable knowledge structure, or ontology, for expressing thematic and textual connections among historical newspaper archives. Even with standards in place, digitization projects pursue differing approaches that pose challenges to integration or particular levels of analysis. In most, for instance, generic identification of items within newspapers has not been pursued. In order to build an ontology, this project will build on knowledge acquired by participating academic partners, such as the project TimeCapsule at Utrecht University, as well as analytical software that has been tested and used by team members, such as viral text analysis. The members of the Finnish team have already worked on text reuse in Finnish newspapers and journals (the COMHIS project), but OcEx offers the possibility to expand this analysis into the study of transnational and transcontinental flows of information. OcEx does not aim to create a totalizing research infrastructure but rather to expose the conditions by which researchers can work across collections, helping guide similar projects in future seeking to bridge national collections. This ontology will be established through comparative investigations of phenomena illustrating textual links: reprinting and topic dissemination. We have divided the tasks into six work packages:

WP1: Management

- create an international network of researchers to discuss issues of using and accessing newspaper repository data and combine expertise toward better development and management of such data;
- assemble a project advisory board, consisting of representatives of public and private data custodians and other critical stakeholders.

WP2: Assessment of Data and Metadata

- investigate and develop classifier models of the visual features of newspaper content and genres;
- create a corpus of annotations on clusters/passages that records relationships among textual versions.

WP3: Creating a Networked Ontology for Research

- create an ontology of genres, forms, and elements of texts to support that annotation;
- select and develop best practices based on available technology (semantic web standard RDF, linked data, SKOS, XML markup standards such as TEI).

WP4: Textual Migration and Viral Texts

- analyze text reuse across archives using statistical language models to detect clusters of reprinted passages;
- perform analyses of aggregate information flows within and across countries, regions, and publications;
- develop adaptive visualization methods for results.

WP5: Conceptual Migration and Translation Shifts

- perform scalable multilingual topic model inference across corpora to discern translations, shared topics, topic shifts, and concept drift within and across languages, using distributional analysis and (hierarchical) polylingual topic models;
- analyze migration and translation of ideas over regional and linguistic borders;
- develop adaptive visualization methods for the results.

WP6: Tools of Delivery/Dissemination

- validation of test results in scholarly contexts/test sessions at academic institutions;
- conduct analysis of the sensitivity of results to the availability of corpora in different languages and levels of access;
- share findings (data structures/availability/compatibility, user experiences) with institutional partners;
- package code, annotated data (where possible), and ontology for public release.

Abstract, DHN 2018
Heini Arjava

Prosodic clashes between music and language – challenges of corpus-use and openness in the study of song texts

In my talk I will discuss the relationship between linguistic and musical rhythm, focusing on the digital methods used in their study and on questions related to open science that arise respectively. This ongoing corpus research examines the correlation between linguistic and musical segment length in songs, more precisely on instances where the language adapts prosodically to the rhythmic frame provided by pre-existing music. The question is to what extent can clashes, by which I mean instances of non-conformity between linguistic and musical segment length**,** be acceptable in song lyrics, and what other prosodic features, such as stress, may influence their occurrence.

Addressing these issues with a corpus-based approach leads to questions of retrieving information from complicated corpora which combine two media (music and language), and the openness and accessibility of music sources. In this abstract I will first describe my research premises in section 1, and discuss in section 2 my corpus methods and their challenges vis-à-vis the digital humanities and open science.

1. Research setting

The natural starting point for the comparison of music and language is the shared interface between western art music and metric language: the isochronic pulse, which consists of rhythmically prominent elements occurring at regular intervals. My study aims to approach this interface by both qualitative and statistical methods.

The study is based on a self-collected song corpus in Finnish, a language where syllable length has a versatile relationship with stress (cf. Hakulinen et al 2004). Primary stress in Finnish is weight-insensitive and always falls on the first syllable of a word, and syllables of any length, long or short, can be stressed or unstressed. Finnish sound segment length is also phonemic, that is, it creates distinctions of meaning. Syllable length in Finnish is therefore of particular interest in a study of musical segment length, because length deviations play an evident role in language perception.

Music and text can be turned into a composition in a number of ways, but my study focuses on the situations in which language is most dependent on music. Usually there are three alternative orders in which music and language can be combined into songs: First, text and music

may be written simultaneously and influence the musical and linguistic choices of the writer at the same time (Language <–> Music). Secondly, text can precede the music, as when composers compose a piece to existing poetry (Language –> Music). And finally, the melody may exist first, as when new versions of songs are created by translating or otherwise rewriting them to familiar tunes (Music –> Language).

My research is concerned with this third relationship, because it poses the strongest constraints on the language user. The language (text) must conform to the music's already existing rhythmic frame that is in many respects inflexible, and in such cases, it is difficult to vary the rhythmic elements of the text, because the musical space restricts the rhythmic tools available for the language user. This in turn may lead to non-neutral linguistic output. Thus, the crucial question arises: How does language adapt its rhythm to music?

A crucial presupposition when problematising the relationship between a musical form and the text written to it is the notion that a song is not poetry per se (I will return to this conception in section 2). The conventions of Western art music allow for a far greater range of length distinctions than language: the syllable lengths usually fall into binary categories (e.g. short and long syllables), whereas in music notes can be elongated infinitely. A translated song in which all rhythmic restrictions come from the music may follow the lines of poetic traditions, but must deviate from them if the limits of space within music do not allow for full flexibility. It is therefore an intermediate form of verbal art.

2. The statistical corpus method, and challenges regarding digital humanities and open science

My corpus contains songs that clearly represent the order of music being created before the text and providing the rhythmic frame of the song. The pilot corpus consists of 15 songs and approximately 1500 prosodically annotated syllables of song texts in Finnish, translated or otherwise adapted, or written to instrumental or traditional music. The genres include chansons, drinking songs, Christmas songs and hymns, which originate from different eras and have originally been written in different languages, including English, French, German, Swedish, and Italian.

I will analyse the data by statistical methods in the R environment. The song texts are annotated in a table syllable by syllable, where one row represents a datapoint (one note/syllable) in one column; and other columns contain metadata about the songs and musical and linguistic prosodic variables, including stress, segment length in musical beats for notes and moras for syllables, and sonority features of the segments.

The moraic length of the syllables will be compared with the length of the respective notes (musical length and stress). The most basic instance of a clash between segment lengths is the instance where a short syllable ((C)V in Finnish) falls on a long note (i.e. a longer note than a basic half-beat). Both theoretical considerations and empirical data will be used in the eventual analysis to determine which length values create the clearest cases of prosodic clashes, and if sonority and stress play a role as well.

The corpus-based approach to language and music raises problematic questions. First of these is, of course, if useful music-linguistic corpora can be found at all at the present. Existent written and spoken corpora of the major European languages contain millions of words, often annotated to a great linguistic detail (cf. Korp of Kielipankki for Finnish (korp.csc.fi), which offers detailed contextual, morphological and syntactic analysis). For music as well, digital music scores can be found "in a huge number" (Ponce de León et al. 2008:560). Corpora of song texts with both linguistic and musical information seem to be more difficult to find.

One problem of music linguistic studies is related to the more restricted openness and shareability of sources than that of written or spoken language. The copyright questions of art are in general a more sensitive issue than for instance those of newspaper articles or internet conversations, and the reluctance of the owners of song texts and melodies may have made it difficult to create open corpora of contemporary music.

But even with ownership problems aside (such as with older or traditional music), building a music-linguistic corpus remains a difficult task to comply. A truly useful corpus of music for linguistic purposes would include metadata and annotation of both media, both language and music. Thus even an automatically analysed metric corpus of poetry, like Anatoli Starostin's Treeton for metrical analysis of Russian poems (Pilshcikov & Starostin 2011) or the rhythmic Metricalizer for determining meter by stress patterns in German poems (Bobenhausen 2011) does not answer to the questions of rhythm of a song text, which exists in a extra-linguistic medium, music, altogether. Vocal music is metrical in the isochronic sense, but it is not metrical in the strict sense of poetic conventions, which are based on linguistic rules. Automated analysis of a song text without its music notation does not tell anything about its real metrical structure.

On a technical level, a set of tools that is necessary for researchers of music are the tools for quick visualization of music passages (notation tools, sound recognition). Such software can be found and used freely in the internet and are useful for depiction purposes. Mining of information from music requires more effort, but has been done in various projects for instance for melody

information retrieval (Ponce de León et al. 2008), or metrical detection of notes (Temperley 2001). But again, these tools seem to rarely combine linguistic and musical meter simultaneously.

By raising these questions I hope to bring attention to the challenges of studying texts in the musical domain, that is, not simply music or poetry separately. The crux of the issue is that for the linguistic analysis of song texts we need actual textual data where the musical domain appears as annotated metadata. Means exist to analyse text automatically, and to analyse musical patterns with sound recognition or otherwise, but to combine the two raises the analysis to a more complicated level. When the issues of effective analysis are solved, it will enable to increase the size and amount of linguistic song corpora.

Literature

Blumenfeld, Lev. 2016. End-weight effects in verse and language. In: *Studia Metrica Poet. Vol. 3.1* pp. 7–32.

Bobenhausen, Klemens. 2011. The Metricalizer – Automated Metrical Markup of German Poetry. In: Küper, C. (ed.), *Current trends in metrical analysis*, pp. 119-131. Frankfurt am Main; New York: Peter Lang.

Hayes, Bruce. 1995. *Metrical Stress Theory: principals and case studies*. Chicago: The University of Chicago Press.

Hakulinen, *et al*. (eds.). 2004. *Iso suomen kielioppi*, pp.44–48. Helsinki: Suomalaisen Kirjallisuuden Seura.

Jeannin, M. 2008. Organizational Structures in Language and Music. In: *The World of Music,50*(1), pp. 5–16.

Kiparsky, Paul. 2006. A modular metrics for folk verse. In: B. Elan Dresher & Nila Friedberg (eds.), *Formal approaches to poetry: recent developments in metrics*, pp.7–52. Berlin: Mouton de Gruyter.

Lerdahl, Fred & Jackendoff, Ray. 1983. *A generative theory of tonal music*. Cambridge (MA): MIT.

Lotz, John. 1960. Metric typology. In: Thomas Sebeok (ed.), *Style in language*. Massachusetts: The M.I.T. Press.

Palmer, Caroline & Kelly, Michael H. 1992. Linguistic Prosody and Musical Meter in Song. *Journal of memory and language* 31, pp. 525–542.

Pilshchikov, Igor & Starostin, Anatoli. 2011. Automated Analysis of Poetic Texts and the Problem of Verse Meter. In: Küper, C. (ed.), *Current trends in metrical analysis*, pp. 133–140. Frankfurt am Main; New York: Peter Lang.

Ponce de León, Pedro J., Iñesta, José M. & Rizo, David. 2008. Mining Digital Music Score Collections: Melody Extraction and Genre Recognition. In: Peng-Yeng Yin (ed.), *Pattern Recognition Techniques, Technology and Applications*, pp. 626–. Vienna: I-Tech.

Temperley, D. 2001. *The Cognition Of Basic Musical Structures*. Cambridge, Mass: MIT Press.

DHN 2018
Roadshow presentation

Facilitating Digital History in Finland

What can we learn from the past?

Presenters: Mats Fridlund, Mila Oiva, Petri Paju

Abstract

The paper discusses the findings of “From Roadmap to Roadshow: A collective demonstration & information project to strengthen Finnish digital history” project. This collaborative research and dissemination project received funding from the Kone Foundation and aims to develop digital history within different history disciplines in Finland. This project and our findings takes as its starting point a survey in 2017 among Finnish historians that identified several critical issues that required further development: better information channels of digital history resources and events, providing relevant education, skills, and teaching by historians, and aiding historians and information technology specialists to meet and collaborate better and more systematically. Many historians also had issues with the concept of digital history and difficulties embracing such an identity.

In order to situate Finnish digital history in the domestic and international contexts, we have studied the 1960s roots of the computational history research in Finland, the current best practice of how to institutionalise and organize digital history internationally. We have visited selected digital humanities centers in Europe and the US identified as having “done something right”. Based on these studies, visits and interviews we will propose steps to be taken for further strengthen the digital history research community in Finland some we already started implementing through creating common meeting places for CS and humanities researchers such as weekly DH lunch meetings. In January-February 2018 we organised a two-week roadshow to six Finnish universities to give workshops teaching practical DH tools and methodologies geared towards historical research as well as to discuss and enquire about the state of the art and future needs of historical researchers.

The presentation discuss what we have learned about the present day conditions of digital history in Finland, how digital humanities is facilitated today in Finland and abroad, and what suggestions we could give for strengthening the conditions for doing digital history research in Finland. We will focus on the critical issues we have identified in surveys and discussions with historians in the Finnish universities, and discuss the possible solutions to these questions found while visiting DH centers.The paper will discuss the following preliminary findings:

- The set-up of different labs in Europe and the US suggests that a laboratory or center supporting an inclusive network are the organizational structures that both supports

stability, provides a space for variation in research, and attracts a critical mass of scholars interested in digital research methods.

- We suggest that there is an importance to informal and low-pressure events and meetings that enable historians to meet computer scientists on and regular and informal basis, such as the weekly “DH pizza” event at the Aalto University.
- As one example of new commonplace practices that maintain interdisciplinary DH network is to develop an “interdisciplinary collaboration pipeline” for joint humanist-computer scientist project collaborations. The importance of this is to make the theory and the essential parts of a collaborative projects visible which appears to be especially valuable for historians who are often trained to work individually, and thus should ease future collaborations.
- Teaching DH methods nationwide with an emphasis of temporary non-regular forms of teaching such as teaching ‘roadshows’ to various universities vs. teaching conducted by staff at a fixed and stable lab.
- We found the roadshow to offer one possible although temporary organisation to be helpful in introducing the digital research methods to all the major Finnish universities, and perhaps as a start toward creating a community of digital historians nationwide. Especially its importance appears to lie in providing an opportunity of a knowledgeable introduction and possibility to informally discuss and pose questions to other historians regarding the promises and pitfalls of digital history.
- We notice that there could be more done to make digital tools for keeping the community together. Historians too could benefit from more robust digital research infrastructures, or from a more organised way of utilizing existing solutions and structures.

These solutions should facilitate digital history research at different levels of abstraction. The provisional conclusion is that digital history in Finland is in a very good state compared to other Scandinavian countries as well as internationally. The presentation will in more detail present and discuss these and other findings from the roadshow concerning previous and present digital history research and organisation of Finnish digital history and digital humanities.

Text Reuse and Eighteenth-Century Histories of England

Ville Vaara, Aleksi Vesanto & Mikko Tolonen, University of Helsinki

Introduction

David Hume's *History* was published after a period of political turbulence in the British Isles. The foundations of the monarchy had been shaken in a series of crisis, with the revolution of 1688 and the ensuing Civil War being the main foci of Hume's *History.* This conflict between Royalists and Parliamentarians, and later Tories and Whigs sets the context in which the works analysed here were both written and received.

But what kind of history is Hume's *History of England*? Is it an impartial account or is it part of a political project? To what extent was it influenced by seventeenth-century Royalist authors? These questions have been asked since the first Stuart volumes were published in the 1750s. The consensus is that Hume's use of Royalist sources left a crucial mark on his historical project.[1] One aim of this paper is to weigh these claims against our evidence about Hume's use of historical sources. To do this we qualified, clustered and compared 129,646 instances text reuse in Hume's *History*. Additionally, we compared Hume's *History of England* with similar undertakings in the eighteenth-century and got an overview of their composition. We aim to extend the discussion on Hume's *History* in the direction of applying computation methods on understanding the writing of history of England in the eighteenth-century as a genre.[2]

This paper contributes to the overall development of Digital Humanities by demonstrating how digital methods can help develop and move forward discussion in an existing research case. We don't limit ourselves to general method development, but rather contribute in the specific discussions on Hume's *History* and study of eighteenth-century histories.

Methods and sources

We are aiming to better understand the composition of Hume's *History* by examining the direct quotes in it based on data in Eighteenth-Century Collections Online (ECCO). It should be noted that ECCO also includes central seventeenth-century histories and other important documents as reprints. Thus, we do not only include eighteenth-century sources, but, for example, works by Clarendon, John Rushworth and other notable seventeenth-century historians. We compare text reuse in Hume's *History* to that in works of Paul de Rapin, William Guthrie and Thomas Carte, all prominent historians at the time. To our knowledge,

---

[1] Royce MacGillivray, 'Hume's "Toryism" and the Sources for his Narrative of the Great Rebellion', *Dalhousie Review*, 56, 1987, pp. 682-6; Laird Okie, 'Ideology and Partiality in David Hume's History of England', *Hume Studies*, vol. 11, 1985, pp. 1-32. See also, Ernest Mossner, "Was Hume a Tory Historian?', *Journal of the History of Ideas*, 2, 1941, pp. 225-236; B. A. Ring, 'David Hume: Historian or Tory Hack?', *North Dakota Quarterly, 1968, pp. 50-59*; Frances Palgrave, 'Hume and his influence upon History' in vol. 9 of *Collected Historical Works*, e.d R. H. Inglis Palgrave, 10 vols. CUP, 1919-22 and Claudia Schmidt, *Reason in history*, 2010.

[2] Previous attempts towards this direction include Karen O'Brien, *Narratives of Enlightenment: Cosmopolitan History from Voltaire to Gibbon*, CUP, 1997.

similar text mining effort has not been previously undertaken in the field of eighteenth-century historiography.

As a starting point for our analysis, we used a dataset of linked text-reuse fragments found in ECCO, constructed with the BLAST -bioanalysis software[3] The basic idea was to create a dataset that identifies similar sequences of characters (from circa 150 to more than 2000 characters each) instead of trying to match individual characters or tokens/words. This helped with the optical character recognition problems that plague ECCO. The methodology has previously been used in matching DNA sequences, where the problem of noisy data is likewise present. We further enriched the results with bibliographical metadata from the English Short Title Catalogue (ESTC). This enriching allows us to compare the publication chronology and locations, and to create rough estimates of first edition publication dates.

There is no ready-to-use gold standard for text reuse cluster detection. Therefore, we compared our clusters with the critical edition of the *Essay Concerning Human Understanding* (EHU) to see if text reuse cases of Hume's *Treatise* in EHU are also identified by our method. The results show that we were able to identify all cases included in EHU except those in footnotes. Because some of the changes that Hume made from the *Treatise* to EHU are not evident, this is a promising result.

Analysis

To give a general overview of Hume's *History* in relation to other works considered, we compared their respective volumes of source text reuse (figure 1). The comparison reveals some fundamental stylistic and structural differences. Hume's and Carte's Histories are composed quite differently from Rapin's and Guthrie's, which have roughly three times more reused fragments: Rapin typically opens a chapter with a long quote from a source document, and moves on to discuss the related historical events. Guthrie writes similarly, quoting long passages from sources of his choice. Hume is different: His quotes are more evenly spread, and a greater proportion of the text seems to be his own original formulations.

[3] Vesanto, Nivala, Salakoski, Salmi & Ginter: A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora. *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24.* May 2017, Gothenburg, Sweden.

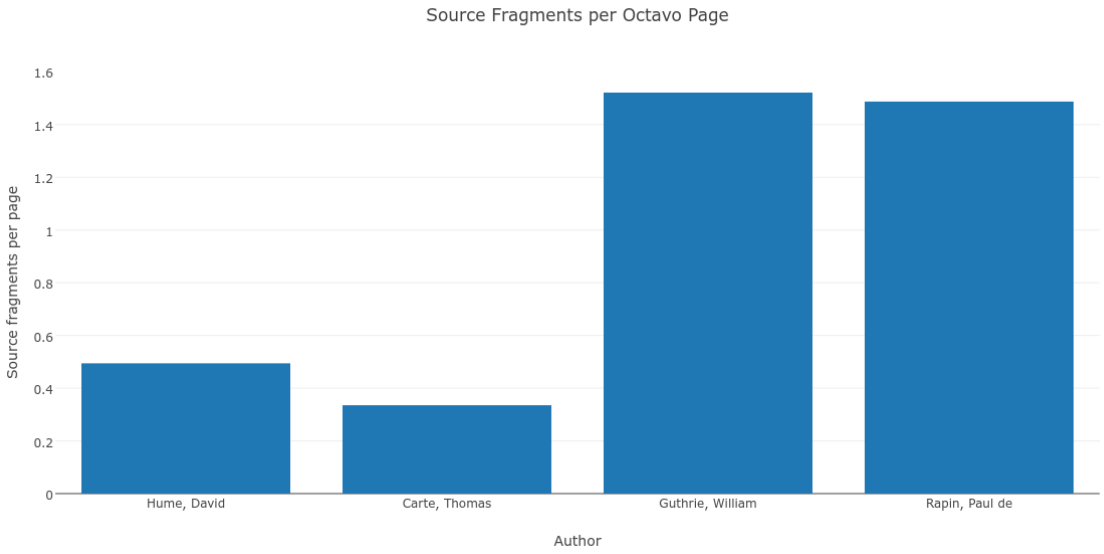


Figure 1.

Change in text reuse in the Histories

All the histories of England considered in our analysis are massive works, comprising of multiple separate volumes. The amount of reused text fragments found in these volumes differs significantly, but the trends are roughly similar. The common overall feature is a rise in the frequency of direct quotes in later volumes.

The increase in text reuse peaks in the volumes covering the reign of Charles I, and the events of the English Civil War, but with respect to both Hume and Rapin (figures 2 & 3), the highest peak is not at the end of Charles' reign, but in the lead up to the confrontation with the parliament. In Guthrie and Carte (figures 4 & 5) the peaks are located in the final volume. Except for Guthrie, all the other historical works considered here have the highest reuse rates located around the period of Charles I's reign that was intensely debated topic among Hume's contemporaries.

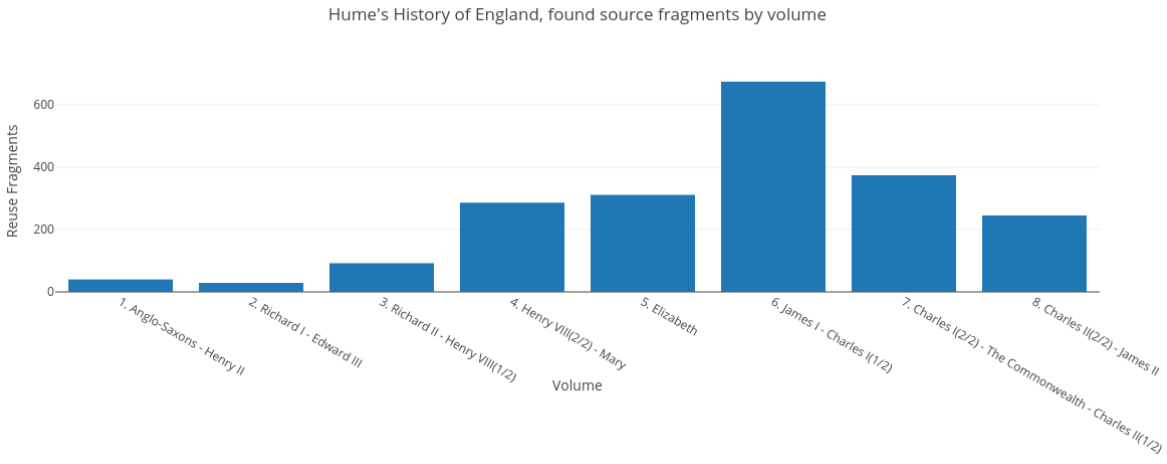


Figure 2.

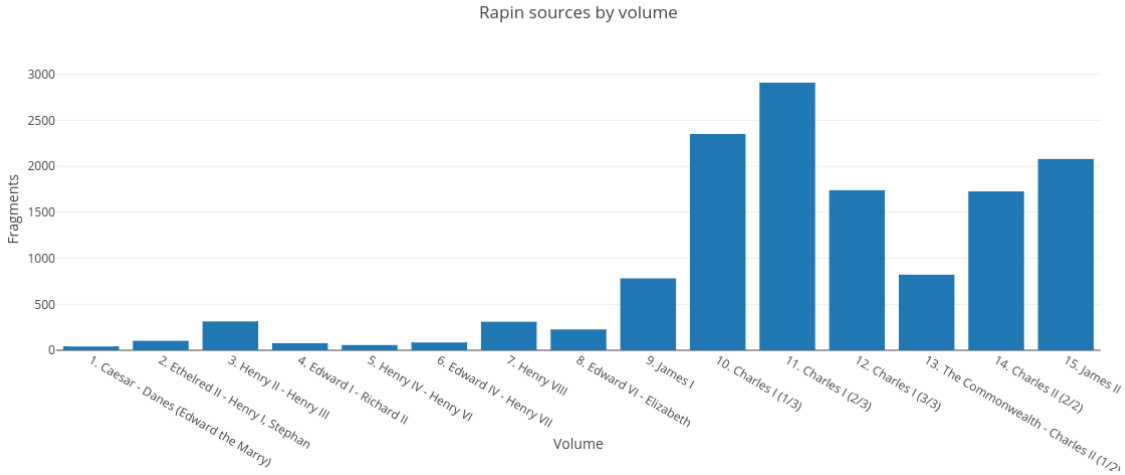
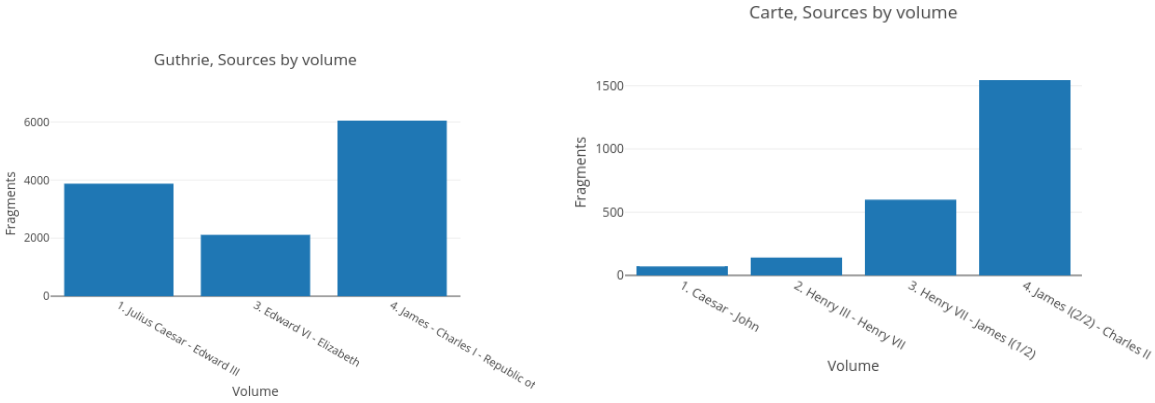


Figure 3.



Figures 4, 5.

We can further break down the the sources of reused text fragments by political affiliation of their authors (figure 6). A significant portion of the detected text reuse cases by Hume link to authors with no strong political leaning in the wider Whig-Tory context. It is obvious that serious antiquary work that is politically neutral forms the main body of seventeenth-century historiography in England. With the later volumes, the amount of text reuses cases tracing back to authors with a political affiliation increases, as might be expected with more heavily politically loaded topics.

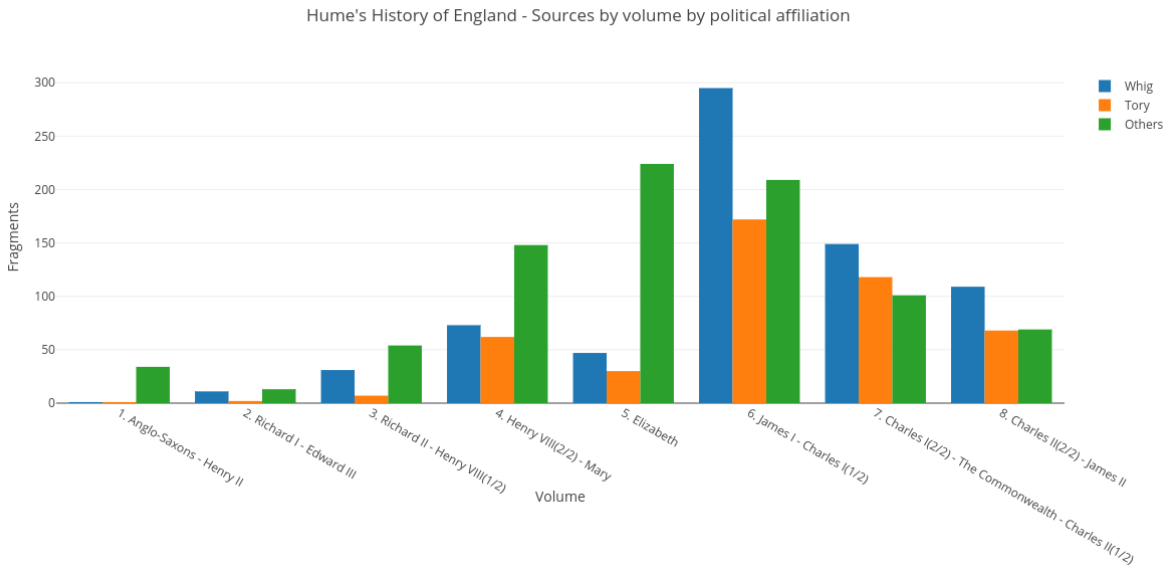


Figure 6.

Charles I execution and Hume's impartiality

A relatively limited list of authors are responsible for majority of the text fragments in Hume's *History*. As one might intuitively expect, the use of particular authors is concentrated in particular chapters. In general, the unevenness in the use of quotes can be seen as more of a norm than an exception.

However, there is at least one central chapter in Hume's Stuart history that breaks this pattern. That is, Chapter LIX - perhaps the most famous chapter in the whole work, covering the execution of Charles I. Nineteenth-century Whig commentators argued, with great enthusiasm, that Hume's use of sources, especially in this particular chapter, and Hume's description of Charles's execution, followed Royalist sources and the Jacobite Thomas Carte in particular. Thus, more carefully balanced use of sources in this particular chapter reveals a clear intention of wanting to be (or appear to be) impartial on this specific topic (figure 7).

Of course, there is John Stuart Mill's claim[4] that Hume only uses Whigs when they support his Royalist bias. In the light of our data, this seems unlikely. If we compare Hume's use of Royalist sources in his treatment of the execution of Charles I to the chapter covering the topic in Carte's work, we note that here Carte relies especially heavily on Royalists, whereas Hume's source use is aligned with his use of Tories elsewhere in the volume.

[4] John Stuart Mill, 'Brodie's History of the British Empire', Robson et al. ed. *Collected works*, vol. 6, pp. 3-58.

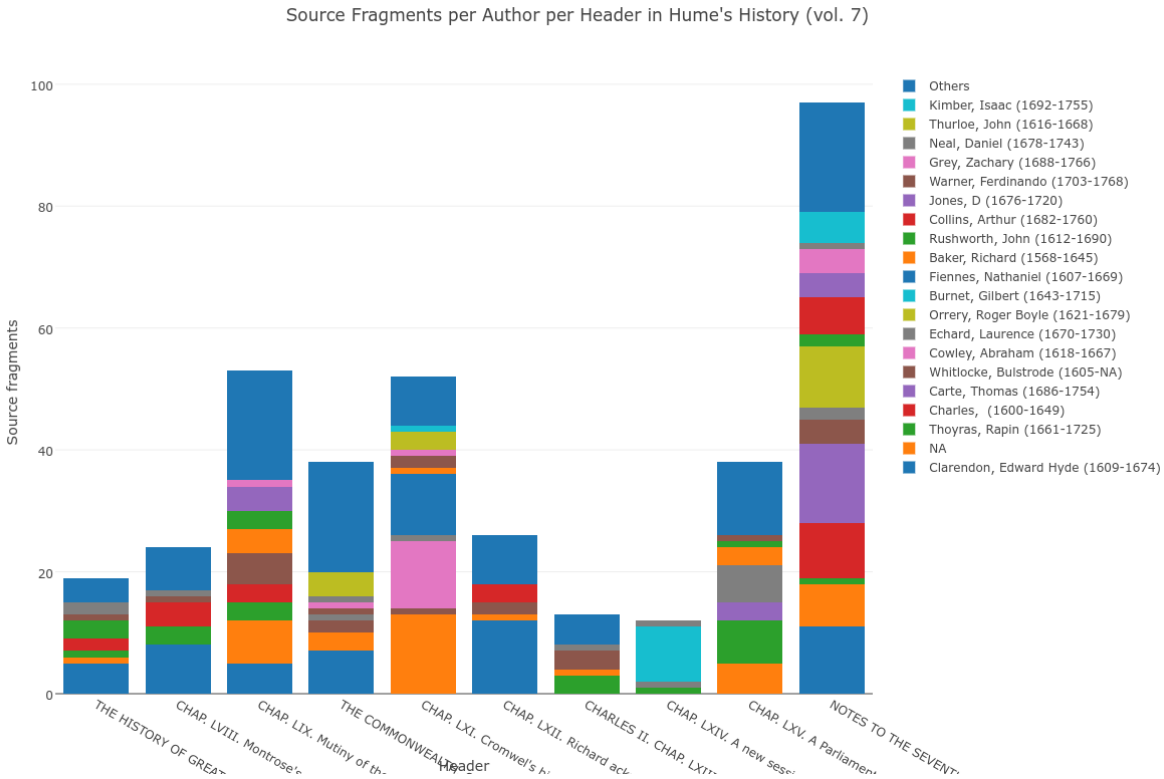


Figure 7.

Hume’s influence on later Histories

A final area of interest in terms of text reuse is what it can tell us about an author’s influence on later writers. The reuse totals of Hume’s *History* in works following its publication are surprisingly evenly spread out over all the volumes (figure 8), and in this respect differ from the other historians considered here (figures 9 - 11). The only exception is the last volume where a drop in the amount of detected reuse fragments differs from the overall image.

Of all the authors only Hume has a high point in reuse at the volumes discussing the Civil War. The reception of Hume’s first Stuart volume, the first published volume of his *History* is well known. It is notable that the next volumes published, that is the following Stuart volumes, possibly written with the angry reception of the first Stuart volume in mind, are the ones that seem to have given rise to least discussion.

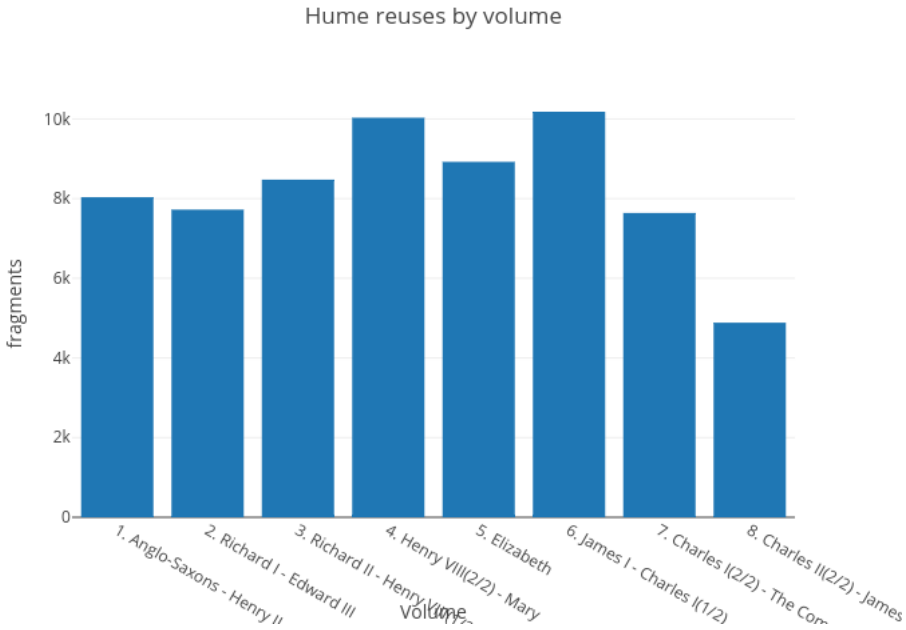


Figure 8.

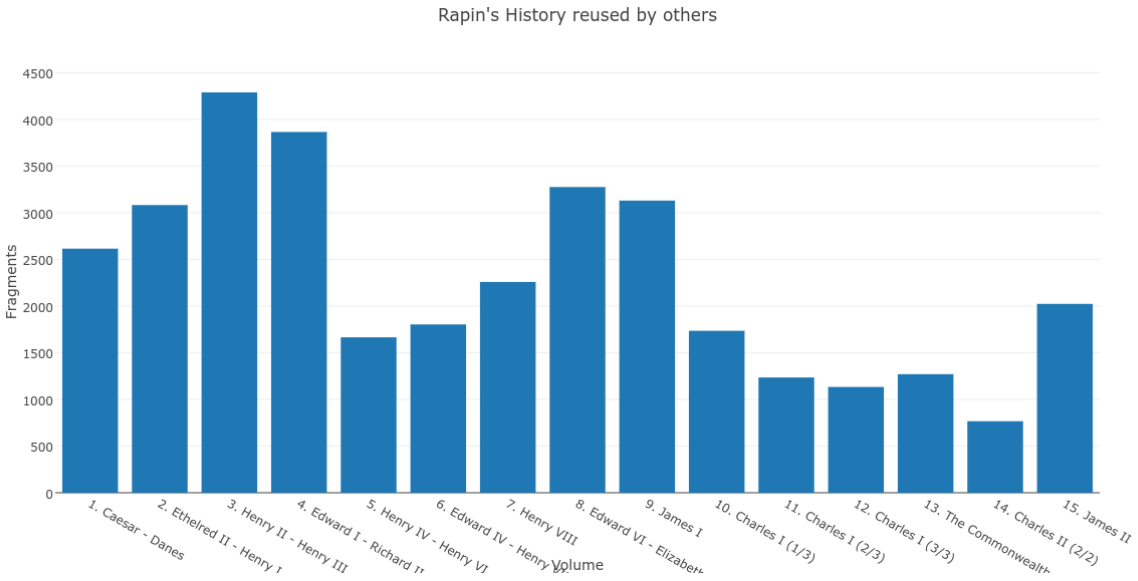
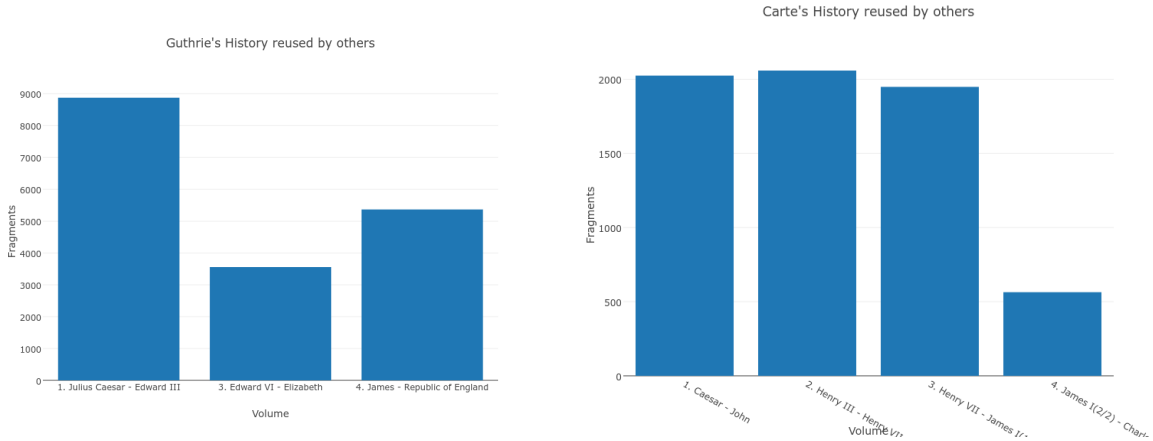


Figure 9.



Figures 10 & 11.

Conclusion

The preliminary results presented above demonstrate how digital methods can open new approaches to historiography. Mapping intertextual connections at a similar volume has not been previously possible, and at best our approach can lead to discovery of overlooked or even unknown influences in literary history. Regarding Hume's *History*, our results reinforce claims that seek to nullify the previously persistent myth of Hume's Toryism, and therefore providing a reason for a closer look at Hume's own ideas about his political impartiality. Additionally, our approach can be further refined and developed towards a tool for mapping out an author's fingerprint of source use and hidden literary influences.

Bibliography

Original sources

Eighteenth-century Collections Online (GALE)
English Short-Title Catalogue (British Library)
Thomas Carte, *General History of England*, 4 vols., 1747-1755.
William Guthrie, *History of Great Britain*, 3 vols., 1744-1751.
David Hume, *History of England*, 8 vols., 1778.
David Hume, *Enquiry concerning Human Understanding*, ed. Tom L. Beauchamp, OUP, 2000.
Paul de Rapin, *History of England*, 15 vols., 1726-32.

Secondary sources

Herbert Butterfield, *The Englishman and his history*, 1944.
John Burrow, *Whigs and Liberals: Continuity and Change in English Political Thought*, 1988.
Duncan Forbes, *Hume's Philosophical Politics*, Cambridge, 1975.
James Harris, *Hume. An intellectual biography*, 2015.
Colin Kidd, *Subverting Scotland's Past. Scottish Whig Historians and the Creation of an Anglo-British Identity 1689–1830*, Cambridge, 1993.
Royce MacGillivray, 'Hume's "Toryism" and the Sources for his Narrative of the Great Rebellion', *Dalhousie Review*, 56, 1987, pp. 682-6.
John Stuart Mill, 'Brodie's History of the British Empire', Robson et al. ed. *Collected works*, vol. 6, pp. 3-58. (http://oll.libertyfund.org/titles/mill-the-collected-works-of-john-stuart-mill-volume-vi-essays-on-england-ireland-and-the-empire)
Ernest Mossner, "Was Hume a Tory Historian?', *Journal of the History of Ideas*, 2, 1941, pp. 225-236.
Karen O'Brien, *Narratives of Enlightenment: Cosmopolitan History from Voltaire to Gibbon*, CUP, 1997.
Laird Okie, 'Ideology and Partiality in David Hume's History of England', *Hume Studies*, vol. 11, 1985, pp. 1-32.
Frances Palgrave, 'Hume and his influence upon History' in vol. 9 of *Collected Historical Works*, e.d R. H. Inglis Palgrave, 10 vols. CUP, 1919-22.
John Pocock, *Barbarism and religion*, vols. 1-2.
B. A. Ring, 'David Hume: Historian or Tory Hack?', *North Dakota Quarterly, 1968, pp. 50-59.*

Claudia Schmidt, *Reason in history*, 2010.

Mark Spencer, ‘David Hume, Philosophical Historian: “contemptible Thief” or “honest and industrious Manufacturer”?, Hume conference, Brown, 2017.

Vesanto, Nivala, Salakoski, Salmi & Ginter: A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora. *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24.* May 2017, Gothenburg, Sweden. (http://www.ep.liu.se/ecp/131/049/ecp17131049.pdf)

Abstract

Johan Jarlbrink & Roger Mähler, Umeå University

Embedded words in the historiography of technology and industry, 1931–2016

Short paper

From 1931 to 2016 The Swedish National Museum of Science and Technology published a yearbook, *Dædalus*. The 86 volumes display a great diversity of industrial heritage and cultures of technology. The first volumes were centered on the heavy industry, such as mining and paper plants located in North and Mid-Sweden. The last volumes were dedicated to technologies and products in people's everyday lives – lipsticks, microwave ovens, and skateboards. During the years *Dædalus* has covered topics reaching from individual inventors to world fairs, media technologies from print to computers, and agricultural developments from ancient farming tools to modern DNA analysis. The yearbook presents the history of industry, technology and science, but can also be read as a historiographical source reflecting shifting approaches to history over an 80-year period. *Dædalus* was recently digitized and can now be analyzed with the help of digital methods.

The aim of this paper is twofold: To explore the possibilities of word embedding models within a humanities framework, and to examine the *Dædalus* yearbook as a historiographical source with such a model. What we will present is work in progress with no definitive findings to show at the time of writing. Yet, we have a general idea of what we would like to accomplish. Analyzing the yearbook as a historiographical source means that we are interested in what kinds of histories it represents, its focus and bias. If words are defined by the distribution of the vocabulary of their contexts we can calculate relations between words and explore fields of related words as well as binary relations in order to analyze their meaning. Simple – and yet fundamental – questions can be asked: What is "technology" in the context of the yearbook? What is "industry"? Of special interest in the case of industrial and technological history are binaries such as rural/urban, man/woman, industry/handicraft, production/consumption, and nature/culture. Which words are close to "man", and which are close to "woman"? Which aspects of the history of technology and industry are related to "production" and which are related to "consumption"?

Word embedding is a comparatively new set of tools and techniques within data science (NLP) with that in common that the words in a vocabulary of a corpus (or several corpora) are assigned numerical representations through some (of a wide variety of different) computation. In most cases, this comes down to not only mapping the words to numerical vectors, but doing so in such a way that the numerical values in the vectors reflect the contextual similarities between words. The computations are based on the distributional hypothesis stemming from Zellig Harris (1954), implicating that "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough, 1965). A predecessor of the method used here is the self-organizing maps, where the contextual roles of words are represented in a

two-dimensional space. (Kohonen, 1995). With word embedding the words are embedded (positioned) in a high-dimensional space, each word represented by a vector in the space i.e. a simple representational model based on linear algebra. The dimension of the space is defined by the size of the vectors and the similarity between words then become a matter of computing the difference between vectors in this space, for instance the difference in (euclidian) distance or difference in direction between the vectors (cosine similarity). Within vector space models the former is the most popular under the assumption that related words tend to have similar directions. The arguably most prominent and popular of these algorithms, and the one that we have used, is the skip-gram model Word2Vec (Mikolov et al, 2013). In short, this model uses a neural network to compute the word vectors as results from training the network to predict the probabilities of all the words in a vocabulary being nearby (as defined by a window size) a certain word in focus.

An early evaluation shows that the model works fine. Standard calculations often used to evaluate the performance and accuracy indicates that we have implemented the model correctly – we can indeed get the correct answers to equations such as "Paris - France + Italy = Rome" (Mikolov et al, 2013). In our case we were looking for "most_similar(positive=['sverige','oslo'], negative=['stockholm'])". And the "most similar" was "norge". We have also explored simple word similarity in order to evaluate the model and get a better understanding of our corpus. What remains to be done is to identify relevant words (or group of words) that can be used when we are examining "topics" and binary dimensions in the corpus. We are also experimenting with different ways to cluster and visualize the data. Although some work remains to be done, we will definitely have results to present at the time of the conference.

Harris, Zellig (1954). Distributional structure. Word, 10(23):146–162.

Kohonen, Teuvo (1995). *Self-organizing maps*. Berlin: Springer

Mikolov, Tomas, Chen, Kai, Corrado, Greg & Dean, Jeffrey (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781

Rubenstein, Herbert & Goodenough, John (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10): 627-633.

Comparing Topic Model Stability Between Finnish, Swedish, English and French

Simon Hengchen[0000−0002−8453−7221], Antti Kanner[0000−0002−0782−1923], Eetu Mäkelä[0000−0002−8366−8414], and Jani Marjanen[0000−0002−3085−4862]

COMHIS
University of Helsinki, Helsinki, Finland,
{simon.hengchen;anti.kanner;eetu.makela;jani.marjanen}@helsinki.fi

1 Abstract

In the recent years, topic modelling has gained increasing attention in the humanities. Unfortunately, little has been done to determine whether the output produced by this range of probabilistic algorithms is revealing signal or is merely producing noise, nor how well it performs on other languages than English. In this paper, we set out to compare topic model stability of parallel corpora in Finnish, Swedish, English, and French, and the effect of lemmatisation on those languages.

2 Context

Topic modelling (TM) is a well-known (following the work of (6; 7)) yet badly understood range of algorithms within the humanities. While a variety of studies within the humanities make use of topic models to answer historical questions (see (3) for a thorough survey), there is no tried and true method that ascertains that the probabilistic algorithm[1] reveals signal and is not merely responding to noise. The rule of thumb is generally that if the results are interesting and reveal a prior intuition by a domain expert, they are considered correct – in the sense that they are a valid entry point into a humongous dataset, and that the proper work of historical research is to be then manually carried out on a subset selected by the algorithm. As pointed out in previous work (10; 4), this, combined with the fact that many humanistic corpora are on the small side, "the threshold for the utility of topic modelling across DH projects is as yet highly unclear." Similarly, topic instability "may lead to research being based on incorrect foundational assumptions regarding the presence or clustering of conceptual fields on a body of work or source material" (4).

Whilst topic modelling techniques are considered language-independent, i.e. "use[] no manually constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like" (8), they encode key

[1] In this work, we choose to use "topic modelling" as a synonym of Latent Dirichlet Allocation (LDA) (2).

assumptions about the statistical properties of language. These assumptions are often developed with English in mind and generalised to other languages without much consideration. We maintain that these algorithms are not language-independent, but language-agnostic at best, and that accounting for discrepancies in how different languages are processed by the same algorithms is necessary basic research for more applied, context-oriented research – especially for the historical development of public discourses in multilingual societies or phenomena where structures of discourse flow over language borders. Indeed, some languages heavily rely on compounding – the creation of a word through the combination of two or more stems – in word formation, while others use determiners to combine simple words. If one considers a white space as the delimitation between words and disregards punctuation (as is usually done with languages making use of the Latin alphabet), the first tendency results in a richer vocabulary than the second, hence influencing TM algorithms that follow the bag-of-words approach. Similarly, differences in grammar – for example, French adjectives must agree in gender and number with the noun they modify, something that does not exist in English – reinforce those discrepancies. Nonetheless, most of this happens in the fuzzy and non-standard preprocessing stage of topic modelling, and the argument could be made that the language neutrality of TM algorithms rests more on it being underspecified with regard to how to pre-process the language. Previous work has tackled this problem: indeed, (5) studies the effect of stemming and concludes that it either helps or hinders the task, depending of the corpus used. More recently, (9) closely look at the effect of lemmatisation on the interpretability of LDA on a morphologically-rich language, Russian.

In this poster, we set out to test topic model stability across languages with regards to corpus size and the effect of lemmatisation. We do so using a custom-made parallel corpus in Finnish, Swedish, English, and French. By selecting those languages, we have a glimpse of how a selection of different languages are processed by TM algorithms. While concentrating on languages spoken in Europe and languages of interest of our collaborative network of linguists, historians and computer scientists, we are still able to examine two crucial variables: one of genetic and one of cultural relatedness. French and Swedish belong to Indo-European (Romance and Germanic branches, respectively) and Finnish is a Finno-Ugrian language. Finnish and Swedish on the other hand share a long history of close language contact and cultural convergence. Because of this, Finnish contains a large number of Swedish loan words, and, perceivably, similar conceptual systems. English (also a language of the Germanic branch, yet highly influenced by French) will serve as a comparison point between all languages, as it is the language that is the most widely used with TM. By doing so, we go further than related work: we study more than one language, and we use lemmatisation rather than stemming – a more "linguistically-aware" choice.

3 Methodology

Building on (4), we use DBpedia (1)'s built-in multilingual graph structure to select entities that exist in all four languages, and extract the content of the `short abstract` entry: generally, a two-to-three-sentence text. Selecting the short abstracts rather than the full content of the corresponding Wikipedia page has the advantage that it "smoothes out" cultural differences: through the reduction of their size to a few sentences, all DBpedia entries have a relatively similar weight in their own respective language corpus as well as across languages.

To explore our hypothesis, we use a parallel corpus of born-digital textual data in Finnish, Swedish, English, and French. Once the corpus, made of 115,547 documents, is constituted, it becomes possible to apply LDA (2) – a parametric topic modelling algorithm that is the most widely used in the humanities.

The resulting models for each language are stored, the corpora reduced in size, LDA is re-applied, the models are stored, corpora re-reduced, etc. Topic models are compared manually between languages at each stage, and programmatically between stages, for all languages. The same workflow is then applied to the lemmatised version of the above-mentioned corpora, and results compared across languages, sizes, and linguistic preprocessing.

Bibliography

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. Springer (2007)

[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)

[3] Brauer, R., Fridlund, M.: Historicizing topic models, a distant reading of topic modeling texts within historical studies. In: International Conference on Cultural Research in the context of "Digital Humanities", St. Petersburg: Russian State Herzen University (2013)

[4] Hengchen, S., O'Connor, A., Munnelly, G., Edmond, J.: Comparing topic model stability across language and size. In: Proceedings of the Japanese Association for Digital Humanities Conference 2016 (2016)

[5] Joachims, T.: Learning to classify text using support vector machines: Methods, theory and algorithms, vol. 186. Kluwer Academic Publishers Norwell (2002)

[6] Jockers, M.L.: Macroanalysis: Digital methods and literary history. University of Illinois Press (2013)

[7] Jockers, M.L., Mimno, D.: Significant themes in 19th-century literature. Poetics 41(6), 750–769 (2013)

[8] Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25(2-3), 259–284 (1998)

[9] May, C., Cotterell, R., Van Durme, B.: Analysis of morphology in topic modeling. arXiv preprint arXiv:1608.03995 (2016)

[10] Munnelly, G., O'Connor, A., Edmond, J., Lawless, S.: Finding meaning in the chaos (2015)

[11] Real, R., Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. Systematic biology 45(3), 380–385 (1996)

Making a bibliography using metadata

The case of the Norwegian photo book

Lars G Johnsen
Arthur Tennøe

In this presentation we will discuss how one might create a bibliography using metadata taken from libraries in conjunction with other sources external to library information. We limit our scope to enumerative bibliographies as opposed to annotated (or descriptive) bibliographies (Carter and Barker (2010).

A bibliography is in general a list of books that satisfy some description or idea, which in this work is specified to as the concept photo book. taken as an artistic book containing photos that are not made for any . Here we are after a special kind of photo book, the book as an art form, distinguishing it from other books containing photos for special (scientific) purposes, while being open towards the possibility that the book might have a narrative, however the story is mainly conveyed via the photographic images.

We believe that the method developed here can be generalized to other topics, especially the use of library metadata, and as such should be of interest to scholars and librarians that work on other topics.

Metadata for books are provided as MARC-posts (see e.g. Library of Congress (2018)) in a national bibliography, which contains information on all books within a national library system. Some of these books are equipped with classificatory information like subject headings and Dewey decimal classification (see e.g. Majumder and Sarma (2007) on WebDewey and references therein), although all books have information about publication date and author, further classifications varies. In the approximately 500 000 books in the Norwegian National Library, a third is equipped with a Dewey decimal classification. However, the number almost doubles to 60 %, if only newer additions (books after 1970) are considered.

For our purposes, we also consulted an external source of Norwegian photographers listed in Wikipedia, who have published photo books among their works. This list is checked against library metadata to extract titles and classification information, like Dewey and subject keywords.

The Dewey number of most interest is the decimal 779 which is about photo books. In contrast to keywords, the Dewey system comes with relatively detailed instructions for applying its codes. Topic words rely more on the subjective interpretation of the librarian in charge of the classification.

Now, any classification (subject heading or Dewey number) and combinations of these will give rise to a list of books: simply select the books that satisfy the combination. The challenge when making a bibliography is finding these combinations. What subject headings will single out a list of photo books?

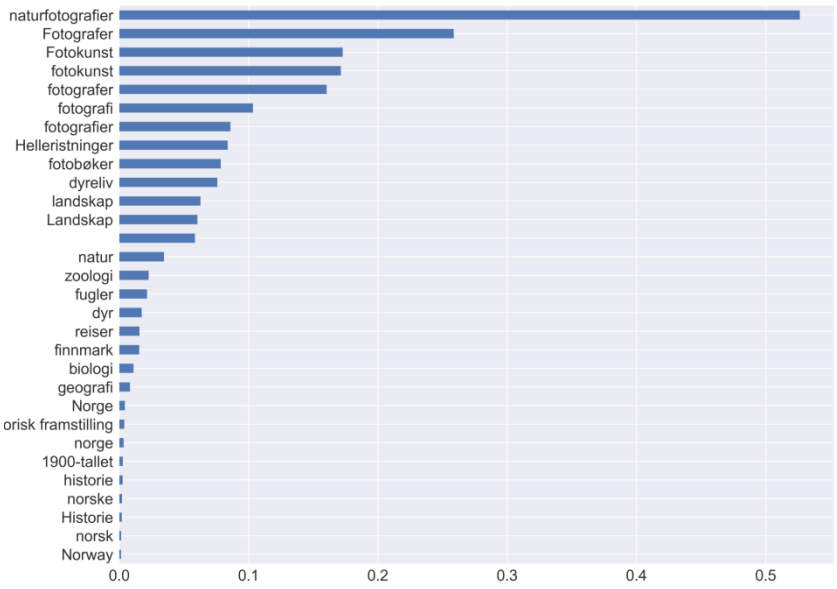
In order to browse the set of topic words (i.e. subject headings) we use the notion of a sister topic. Every book is classified in more than one way, while the Dewey number is unique (up to which library responsible for classification), there are a number of other topics occurring together with a particular topic keyword. The co-occurrence and the frequency will give hints as to which other topic words that may be worth pursuing. Some may be relevant while some others are not.

In the following figure we show a heat map illustrating the topics for keywords relate to photography. This particular table is sorted according to the keyword “Fotokunst” (photographic art), where the capitalized key words signalizes that it belongs to a restricted keyword thesaurus within the Norwegian library system, while unrestricted keywords are used with lower case initial letters. The heat map highlights keywords within each column and the value in each cell shows how many books the row keyword has in common with the columnar keyword. The cell with identical row and column keyword contains the number of titles for that particular keyword.

	Fotohistorie	Fotokunst	foto	fotobok	fotobøker	fotografi	fotografier	fotohistorie	fotokunst	kunstfotografi
Fotokunst	0	280	5	4	31	65	68	8	51	2
Norge	3	110	15	10	33	33	98	17	29	3
Fotografer	2	86	1	0	3	34	30	7	9	2
fotografier	1	68	16	2	52	49	448	9	29	0
fotografi	2	65	10	0	4	292	49	9	52	4
fotokunst	0	51	3	0	7	52	29	0	111	1
norge	2	33	8	10	37	69	111	10	13	2
fotobøker	0	31	1	25	435	4	52	2	7	0
fotografer	0	28	1	0	0	45	25	5	43	0
natur	0	28	1	2	46	24	18	1	5	0
Motiver (Bildekunst)	0	23	0	0	1	10	7	0	8	0
Fotografisamlinger	1	22	0	0	1	4	5	1	9	0
Lofoten	0	20	0	0	7	16	1	0	2	0
Fotografi	3	19	8	0	1	44	28	10	3	1
Fotografering	0	18	8	2	1	31	20	3	3	1
dikt	0	18	0	1	3	18	9	1	4	0
Portrett	0	18	1	0	0	3	5	1	3	0

Such tables can be used to study the relationship between keywords, and how they can be used to refine sets of books. For instance

For the list of Norwegian photographers, we also conducted a study of the topic words associated with their books. These are shown in the following graph, which is adjusted so that typical keywords appear first. Note that while nature photography (naturfotografi) is at the top on this list, photographic art is high (fotokuns).



Using the above information, together with Dewey classifications and keywords we were able to construct a list of about 100 titles.

In addition to the above information we plan also to use the full text information. Photographic books tend to have a small text to photo ratio, which can be used in order to get candidates for books, particularly for those that are not classified with respect to Dewey or topic word. This will perhaps also have a wider interest for projects trying to build bibliographies in general.

References

John Carter; Nicolas Barker (2016). "Bibliography". ABC for Book Collectors (9th ed.). Oak Knoll Press and British Library

Library of Congress (2018) (https://www.loc.gov/marc/)

Majumder, Apurba Jyoti; Gautam Sarma. "Webdewey: The Dewey Decimal Classification in The Web" INFLIBNET Centre, Ahmedabad, Planner 2007. As PDF: http://ir.inflibnet.ac.in:8080/ir/ViewerJS/#../bitstream/1944/1047/1/16.pdf

Designing a Generic Platform for Digital Edition Publishing

Niklas Liljestrand
Svenska litteratursällskapet i Finland

This presentation describes the technical design for streamlining work with publishing Digital Editions on the web. The goal of the project is to provide a platform for scholars working with Digital Editions to independently create, edit, and publish their work. The platform is to be generic, but with set rules of conduct and processes, providing rich documentation of use.

The work on the platform started during 2016, with a rebuild of the website for Zacharias Topelius Skrifter for the mobile web (presented during DHN 2017, http://dhn2017.eu/abstracts/#_Toc475332550). The work continues with building the responsive site to be easily customizable and suite the different Editions needs.

The platform will consist of several independent tools, such as tools for publishing, version comparison, editing, and tagging XML TEI formatted documents. Many of the tools are already available today, but they are heavily dependent on customization for each new edition and MS Windows only. For the existing tools, the project aims to combine, simplify and make the tools platform independent.

The project will be completed within 2018 and the aim is to publish all tools and documentation as open-source.

Timo Korkiakangas, University of Oslo, 3 Jan 2018

Network visualization for historical corpus linguistics: externally-defined variables as node attributes

In my poster presentation, I will explore whether and how network visualization can benefit philological and historical-linguistic research. This will be implemented by examining the usability of network visualization for the study of early medieval Latin scribes' language competences. Thus, the scope is mainly methodological, but the proposed methodological choices will be illustrated by applying them to a real data set. Four linguistic variables extracted corpus-linguistically from a treebank will be examined: spelling correctness, classical Latin prepositions, genitive plural form, and <ae> diphthong. All the four are continuous, which is typical of linguistic variables. The variables represent different domains of language competence of the scribes who learnt written Latin practically as a second-language by that time (Korkiakangas 2017, Korkiakangas & Lassila [submitted]). Even more linguistic features will be included in the analysis if my ongoing project proceeds as planned.

Thus, the primary objective of the study is to find out whether the network visualization approach has demonstrable advantages compared to ordinary cross-tabulations as far as support to philological and historical-linguistic argumentation is concerned. The main means of visualization will be the gradient colour palette in Gephi, a widely used open-source network analysis and visualization software package. As an inevitable part of the described enterprise, it is necessary to clarify the scientific premises for the use of network environment to display externally-defined values of linguistic variables. It is obvious that in order to be utilized for research purposes, network visualization must be as objective and replicable as possible.

By way of definition, I emphasize that the proposed study will not deal with linguistic networks proper, i.e. networks which are directly induced or synthesized from a linguistic data set and represent abstract relations between linguistic units (Araújo & Banisch 2016). Consequently, no network metric will be calculated, even though that might be interesting as such. What will be visualized are the distributions of linguistic variables that do not arise from the network itself, but are derived externally from a medium-sized treebank by exploiting its lemmatic, morphological, and, hopefully, also syntactic annotation layers. These linguistic variables will be visualized as attributes of the nodes in the trimodal "social" network which consists of the documents, persons, and places that underlie the treebank (cf. Bergs 2005). These documents, persons, and places are encoded as metadata in the treebank. The nodes are connected to each other by unweighted edges. The number of document nodes is 1,040, scribe nodes 220, and writing place nodes 84. In most cases, the definition of the 220 writer nodes is straightforward, given that the scribes scrupulously signed what they wrote, with the exception of eight documents. The place nodes are more challenging. Although 78% of the documents has been written in the city of Lucca, the disambiguation and re-grouping of small localities of which little is known was time-consuming and the results not always fully satisfying. The nodes will be set on the map background by utilizing Gephi's Geo Layout and Force Atlas 2 algorithms.

The linguistic features that will be visualized reflect the language change that took place in late Latin and early medieval Latin, roughly the 3rd to 9th centuries AD (Adams 2013). The features are operationalized as variables which quantify the variation of those features in the treebank. This quantification is based on the numerical output of a plethora of corpus-linguistic queries which extract from the treebank all constructions or forms that meet the relevant criteria. The variables indicate the relative frequency of the examined features in each document, scribe, and writing place. For the scribes and writing places, the percentages are calculated by counting the occurrences within all the documents written by that scribe or in that place, respectively.

The resulting linguistic variables are continuous, hence the practicality of the gradient colouring. In order to ground colouring in the statistical dispersion of the variable values and to conserve maximal visual effect, I customize the Gephi default red-yellow-blue palette so that the maximal yellow, which stands for the middle of the colour scale, marks the mean of the distribution of each variable. Likewise, the thresholds of the maximal red and maximal blue are set equally far from the mean. I chose that distance to be two standard deviations away from the mean. In this way, only around 2.5% of the nodes with the lowest and highest values at both ends of the distribution are maximally saturated with red and blue while the rest, around 95%, of the nodes feature a gradient colour, including the maximal yellow in between. Following this rule, I will illustrate the variables both separately and as a sum variable. The images will be available in the poster. The sum variable will be calculated by aggregating the standardized simple variables (cf. Korkiakangas & Lassila [submitted]).

The below image illustrates, by way of an example, the distribution of the spelling correctness variable within the LLCT2 network. The spelling correctness indicates the percentage of characters which are spelled according to the Classical Latin spelling in relation to all the characters of a document. For example, the word form *atmodo* differs from the classical standard form *admodum* "greatly" by three characters. The correct characters are four and, thus, the spelling correctness percentage of the form *atmodo* is 57 (i.e. 4 in 7). Technically, the number of misspelled characters is obtained by calculating the Levenshtein edit distance between each word attested in LLCT2 and the normalized, standard version of that word (Korkiakangas 2017). The red colour stands for a low spelling correctness percentage and blue for a high percentage. The interactive version of the graph with node labels can be consulted at http://bit.ly/2Abk3Bv. That version is realized by SigmaJS tools (http://sigmajs.org/).

The graph shows that the spelling correctness values above the mean are mostly concentrated around one spot, which represents the city of Lucca. Even more importantly, all the substantial blue clusters of high-value documents are written in Lucca, whereas the blue location nodes outside Lucca are due to sporadic high-value documents (and scribes). Conversely, most of the red and reddish low-value nodes are situated outside Lucca, e.g. in Pisa and in peripheral southern and south-western localities. All this elicits the conclusion that classical spelling was cherished primarily in Lucca, the administrative and cultural centre of Tuscia. In sum, the applied distributionally-based principle of gradient colouring seems to be suitable at least for variables which are not too badly divergent from normal distribution. The result is a graph with easily observable colour patterns that are, at the same time, grounded in statistical reality (Korkiakangas & Lassila [submitted]).

The preliminary conclusions also include the observation that network visualization, as such, is not a sufficient basis for philological or historical-linguistic argumentation, but if used along with statistical approach, it can support argumentation by drawing attention to unexpected patterns and – on the other hand – to irregularities. However, it is the geographical layout of the graphs that gives the most of the surplus in regard to traditional approaches: it helps in perceiving patterns that would have otherwise failed to be noticed.

The treebank on which the analyses are based is the Late Latin Charter Treebank (version 2, LLCT2), which consists of 1,040 early medieval Latin documentary texts (c. 480,000 words). The documents have been written in historical Tuscia (Tuscany), Italy, between AD 714 and 897, and are mainly sale or purchase contracts or donations, accompanied by a few judgements as well as lists and memoranda. LLCT2 is still under construction and only the first half of it is already provided with the syntactically annotated layer, thus making it a treebank proper (i.e. LLCT, version 1). The lemmatization and morphological annotation style are based on the Ancient Greek and Latin Dependency Treebank (AGLDT) style which can be deduced from the *Guidelines for the Syntactic Annotation of Latin Treebanks* (Bamman & al. 2007). Korkiakangas & Passarotti (2011) define a number of additions and modifications to these general guidelines which are designed for Classical Latin. For a more detailed description of the LLCT2 and the underlying text editions, see Korkiakangas (2017). Documents are privileged material for examining the spoken/written interface of early medieval Latin, in which the distance between the spoken and written codes had grown considerable by the Late Antiquity. The LLCT2 documents have precise dating and location metadata and they survive as originals.

Bibliography

Adams J.N. Social variation and the Latin language. Cambridge University Press (Cambridge), 2013.

Araújo T. and Banisch S. Multidimensional Analysis of Linguistic Networks. Mehler A., Lücking A., Banisch S., Blanchard P. and Job, B. (eds) *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*. Springer (Berlin, Heidelberg), 2016, 107-131.

Bamman D., Passarotti M., Crane G. and Raynaud S. Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3), 2007 http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf.

Barzel B. and Barabási A.-L. Universality in network dynamics. *Nature Physics*. 2013;9:673-681.

Bergs A. *Social Networks and Historical Sociolinguistics: Studies in Morphosyntactic Variation in the Paston Letters*. Walter de Gruyter (Berlin), 2005.

Ferrer i Cancho R. Network theory. Hogan P.C. (ed.) *The Cambridge Encyclopedia of the Language Sciences*. Cambridge University Press (Cambridge), 2010, 555–557.

Korkiakangas T. Spelling Variation in Historical Text Corpora: The Case of Early Medieval Documentary Latin. *Digital Scholarship in the Humanities*, 2017. https://doi.org/10.1093/llc/fqx061

Korkiakangas T. and Lassila M. Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material. Mambrini F., Sporleder C. and Passarotti M. (eds) *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-3), Sofia, December 13, 2013. Bulgarian Academy of Sciences (Sofia), 2013, 61-72.

Korkiakangas T. and Lassila M. Visualizing linguistic variation in a network of Latin documents and scribes. Manuscript submitted to *Journal of Data Mining and Digital Humanities*.

Korkiakangas T. and Passarotti M. Challenges in Annotating Medieval Latin Charters. Journal of Language Technology and Computational Linguistics. 2011;26,2:103-114.

Abstract for long format presentation
DHN2018

Heidi Karlsen, University of Oslo
Ph.D. Candidate in literature,
Cand.philol. in philosophy

Interdisciplinary advancement through the unexpected: Mapping gender discourses in Norway (1840-1913) with *Bokhylla*

This presentation discusses challenges related to sub-corpus topic modeling in the study of gender discourses in Norway from 1840 till 1913 and the role of interdisciplinary collaboration in this process. Through collaboration with the Norwegian National Library, data-mining techniques are used in order to retrieve data from the digital source, *Bokhylla* [«the Digital Bookshelf»], for the analysis of women's «place» in society and the impact of women writers on this discourse. My project is part of the research project «Data-mining the Digital Bookshelf», based at the University of Oslo.

1913, the closing year of the period I study, is the year of women's suffrage in Norway. I study the impact women writers had on the debate in Norway regarding women's «place» in society, during the approximately 60 years before women were granted the right to vote. A central hypothesis for my research is that women writers in the period had an underestimated impact on gender discourses, especially in defining and loading key words with meaning (drawing on mainly Norman Fairclough's theoretical framework for discourse analysis). In this presentation, I examine a selection of Swedish writer Fredrika Bremer's texts, and their impact on gender discourses in Norway.

The Norwegian National Library's Digital Bookshelf, is the main source for the historical documents I use in this project. The Digital Bookshelf includes a vast amount of text published in Norway over several centuries, text of a great variety of genres, and thus offers unique access to our cultural heritage. Sub-corpus topic modeling (STM) is the main tool that has been used to process the Digital Bookshelf texts for this analysis. A selection of Bremer's work has been assembled into a sub-corpus. Topics have then been generated from this corpus and then applied to the full Digital Bookshelf corpus. During the process, the collaboration with the National Library has been essential in order to overcome technical challenges. I will reflect upon this collaboration in my presentation. As the data are retrieved, then analyzed by me as a humanities scholar, and weaknesses in the data are detected, the programmer, at the National Library assisting us on the project, presents, modifies and develops tools in order to meet our challenges. These tools might in turn represent additional possibilities beyond what they were proposed for. New ideas in my research design may emerge as a result. Concurrently, the algorithms created at such a stage in the process, might successively be useful for scholars in completely different research projects. I will mention a few examples of such mutually productive collaborations, and briefly reflect upon how these issues are related to questions regarding open science.

In this STM process, several challenges have emerged along the way, mostly related to OCR errors. Some illustrative examples of passages with such errors will be presented for the purpose of discussing the measures undertaken to face the problems they give rise to, but also for demonstrating the unexpected progress stemming from these «defective» data. The topics used as a «trawl line»[1], in the initial phase of this study, produced few results. Our first attempt to get more results was to revise down the required Jaccard similarity value[2]. This entails that the

---

[1] My description of the STM process, with the use of tropes such as «trawl line» is inspired by Peter Leonard and Timothy R. Tangherlini (2013): "Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research" in Poetics. 41, 725-749

[2] The Jaccard index is taken into account in the ranging of the scores. The best hit passage for a topic, the one with highest score, will be the one with highest relative similarity to the other captured passages, in terms of concentration of topic words in the passage. The parameterized value of the required Jaccard similarity defines the score a passage must receive in order to be included in the list of captured passages from the «great unread».

Heidi Karlsen, University of Oslo
Ph.D. Candidate in literature,
Cand.philol. in philosophy

quantity of a topic that had to be identified in a passage in order for it to qualify as a hit, is lowered. As this required topic quantity was lowered, a great number of results were obtained. The obvious weakness of these results, however, is that the rather low required topic match, or relatively low value of the required Jaccard similarity, does not allow us to affirm a connection between these passages and Bremer's text. Nevertheless, the results have still been useful, for two reasons. Some of the data have proven to be valuable sources for the mapping of gender discourses, although not indicating anything regarding women writers' impact on them. Moreover, these passages have served to illustrate many of the varieties of OCR errors that my topic words give rise to in text from the period I study (frequently in Gothic typeface). This discovery has then been used to improve the topics, which takes us to the next step in the process.

In certain documents one and the same word in the original text has, in the scanning of the document, given rise to up to three different examples of OCR errors[3]. This discovery indicates the risk of missing out on potentially relevant documents in the «great unread»[4]. If only the correct spelling of the words is included in the topics, potentially valuable documents with our topic words in them, bizarrely spelled because of errors in the scanning, might go unnoticed. In an attempt to meet this challenge I have manually added to the topic the different versions of the words that the OCR errors have given rise to (for instance for the word «kjærlighed» [love] «kjaerlighed», «kjcerlighed», «kjcrrlighed»). We cannot in that case, when we run the topic model, require a one hundred percent topic match, perhaps not even 2/3, as all these OCR errors of the same word are highly unlikely to take place in all potential matches[5]. Such extensions of the topics, condition in other words our parameterization of the algorithm: the required value of Jaccard similarity for a passage to be captured has to be revised fairly down. The inconvenience of this approach, however, is the possible high number of captured passages that are exaggeratedly (for our purpose) saturated with the semantic unit in question. Furthermore, if we add to this the different versions of a lexeme and its semantic relatives that in some cases are included in the topic, such as «kvinde», «kvinder», «kvindelig», kvindelighed» [woman, women, feminine, femininity], the topic in question might catch an even larger number of passages with a density of this specific semantic unity with its variations; this is an amount that is not proportional to the overall variety of the topic in question.

This takes us back to the question of what we program the "trawl line" to "require" in order for a passage in the target corpus to qualify as a hit, and as well to how the scores are ranged. How many of the words in the topic, and to what extent do several occurrences of *one* of the topic's words, i.e., five occurrences of "woman" in one paragraph interest us? The parameter can be set to range scores in function of the occurrences of the different words forming the topic, meaning that the score for a topic in a captured passage is proportional to the heterogeneity of the occurrences of the topic's words, not only the quantity. However, in some cases we might, as mentioned, have a topic comprising several forms of the same lexeme and its semantic relatives and, as described, several versions of the same word due to OCR errors. How can the topic model be programmed in order to take into account such occurrences in the search for matching

---

[3] Some related challenges were described by Kimmo Kettunen and Teemu Ruokolainen in their presentation, «Tagging Named Entities in 19th century Finnish Newspaper Material with a Variety of Tools» at DHN2017.

[4] Franco Moretti (2000) (drawing on Margareth Cohen) calls the enormous amount of works that exist in the world for «the great unread» (limited to *Bokhylla's* content in the context of my project) in: «Conjectures of World Literature» in New Left Review. 1, 54-68.

[5] As an alternative to include in the topic all detected spelling variations, due to OCR errors, of the topic words, we will experiment with taking into account the Levenshtein distance when programming the «trawl line». In that case it is not identity between a topic word and a word in a passage in the great unread that matters, but the distance between two words, the minimum number of single-character edits required to change one word into the other, for instance «kuinde»-> «kvinde».

Heidi Karlsen, University of Oslo
Ph.D. Candidate in literature,
Cand.philol. in philosophy

passages? In order to meet this challenge, a «hyperlexeme sensitive» algorithm has been created[6]. This means that the topic model is parameterized to count the lexeme frequency in a passage. It will also range the scores in function of the occurrence of the hyperlexeme, and not treat occurrences of different forms of one lexeme equally to the ones of more semantically heterogenous word-units in the topic. Furthermore, and this is the point to be stressed, this algorithm is programmed to treat miss-spelling of words, due to OCR errors, as if they were different versions of the same hyperlexeme.

The adjustments of the value of the Jaccard similarity and the hyperlexeme parameterization are thus measures conducted in order to compensate for the mentioned inconveniences, and improve and refine the topic model. I will show examples that compare the before and after these parameters were used, in order to discuss how much closer we have got to be able to establish actual links between the sub-corpus, and passages the topics have captured in the target corpus. All the technical concepts will be defined and briefly explained as I get to them in the presentation. The genesis of these measures, tools and ideas at crucial moments in the process, taking place as a result of unexpected findings and interdisciplinary collaboration, will be elaborated on in my presentation, as well as the potential this might offer for new research.

[6] By the term «hyperlexeme» we understand a collection of graphemic occurences of a lexeme, including spelling errors and semantically related forms.

Our data futures: Towards *non-data-centric* data activism

Minna Ruckenstein, University of Helsinki & Tuukka Lehtiniemi, Aalto University

Key words: datafication, social imaginary, data activism, digital rights, MyData

The social science debate that attends to the exploitative forces of the quantification of aspects of life previously experienced in qualitative form, recognising the ubiquitous forms of datafied power and domination, is by now an established perspective to question datafication and algorithmic control (Ruckenstein and Schüll, 2017). Drawing from the critical political economy and neo-Foucauldian analyses researchers have explored the effects of the datafication (Mayer-Schönberger and Cukier. 2013; Van Dijck, 2014) on the economy, public life, and self-understanding. Studies alert us to threats to privacy posed by "dataveillance" (Raley, 2012; Van Dijck, 2014), forms of surveillance distributed across multiple interested parties, including government agencies, insurers, operators, data aggregators, analytics companies, and individuals who provide the information either knowingly or unintentionally when going online, using self-tracking devices, loyalty programs, and credit cards. The "data traces" add to the data accumulated in databases and personal data – any data related to a person or resulting from actions by a person – becomes utilized for business and societal purposes in an increasingly systematic matter (Van Dijck and Poell, 2016; Zuboff, 2015).

In this presentation, we take an "activist stance", aiming to contribute to the existing criticism of datafication with a more participatory and collaborative approach offered by "data activism" (Baack 2015; Milan and van der Velden, 2016). The various data-driven initiatives currently under development suggest that the problematic aspects of datafication, including the tension between data openness and data ownership (Neff, 2013), the asymmetries in terms of data usage and distribution (Wilbanks and Topol, 2016; Kish and Topol, 2015) and the inadequacy of existing informed consent and privacy protections (Sharon, 2016) are by now not only well recognized, but they are generating new forms of civic and political engagement and activism. This calls for more debate on what new data initiatives and forms of data activism are, and how scholars in the humanities and social science communities can assess them.

By relying on the approaches developed within the field of Techno-Anthropology (Børsen and Botin, 2013; Ruckenstein and Pantzar, 2015), seeking to translate and mediate knowledge concerning complex technoscientific projects and aims, we positioned ourselves as "outside insiders" with regard to a data-centric initiative called MyData. In 2014, we became observers and participants of MyData, promoting the understanding that people benefit when they can control data gathering and analysis by public organizations and businesses and become more active data citizens and consumers. The high-level MyData vision, described in 'the MyData white paper' written primarily by researchers at the Helsinki Institute for Information Technology and the Tampere University of Technology (Poikola et al., 2015), outlines an alternative future that transforms the 'organisation-centric system' into 'a human-centric system' that treats personal data as a resource that the individual can access, control, benefit and learn from[1].

[1] We refer here to what the MyData activists call 'the MyData white paper', the English-language document (Poikola et al. 2015) which is a summary of a Finnish study commissioned by the Ministry of Transport and Communication in (Poikola et al 2014). The aim behind the study was to promote public discussion of the model's potential and impact with respect to handling of personal data. The Finnish-language version is more comprehensive in its outlining of the MyData approach, and we rely also on that version in our discussion.

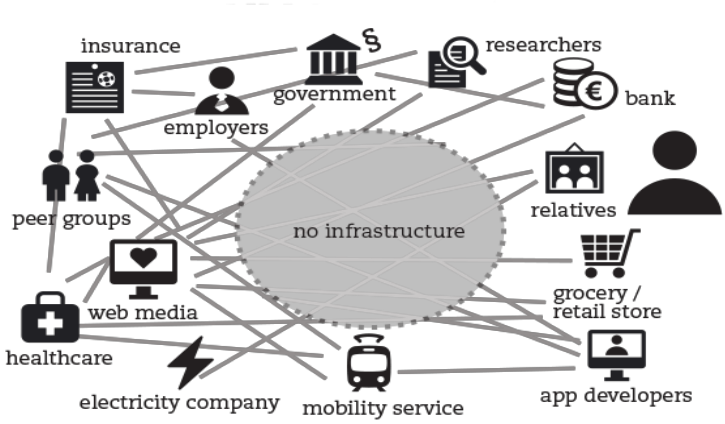


Figure 1 illustrates how the MyData developers perceive their vision compared with the current situation.

The presentation discusses "our" data activist stance and the activism of technology developers, promoting and relying on two different kinds of "social imaginaries" (Taylor, 2004). The notion of the social imaginary, offered by the political philosopher Charles Taylor (2004), aids in the exploration of how people make sense of society's practices, imagine their social existence, and deal with "the expectations that are normally met, and the deeper normative notions and images that underlie these expectations" (Taylor, 2004: 106). By relying on the concept of social imaginary, we open a perspective onto data activism that highlights contested social expectations, along with their ideological and political underpinnings. As we explain in our forthcoming paper, the aim of this exercise is to compare different forms of data activism in order to clarify the kinds of political and social alternatives they offer. Current data-driven initiatives often proceed with a social imaginary

that treats data arrangements as solutions to, or corrective measures for, unsatisfactory developments. They advance the logic of an innovation culture reliant on the development of new technology structures and computationally intensive tools, thereby encompassing an engineering attitude that does not question the power of technological innovation to provide societal solutions or, more broadly, the role of datafication in societal development (Baack, 2015).
Instead, the goal of our activist stance is to introduce other social aims and expectations; informed by the critical stance representative of social scientific inquiry, it questions the optimistic and future-oriented social imaginary of technology developers; indeed, as we will demonstrate, the position we take is incompatible with the engineering attitude in a profound sense. In order to craft a narrative about the MyData initiative that aligns with our social imaginary, we wanted to push the conversation beyond the usual technological, legal, and policy frameworks, and suggest that, with its techno-optimism, current MyData work might in fact weaken data activism and public support for it (Kennedy, forthcoming). We turned to literary and scholarly sources with the aim of opening a critical, but hopefully also a productive conversation about MyData in order to offer ideas of how to promote socially more robust data activism. A seminal text that provides insight into the MyData initiative is the *Autonomous Technology – Technics-out-of-Control as a Theme in Political Thought* (1978) by Langdon Winner. Winner perceives the relationship between human and technology in terms of Kantian autonomy: via analysis of interrelations of independence and dependence. The core ideas of the MyData vision have particular resonance with the way Winner (1978) considers "reverse adaptation", wherein the human adapts to the power of the system and not the other way around.

In our presentation, we first describe the MyData vision, as it has been presented by the activists, and then situate it in the framework of technology critique and current critique of digital culture and economy. Here, we demonstrate that the outside position can, in fact, resource a re-articulation of data activism. After this, we detail some further developments in the MyData scene and possibilities that have opened for dialogue and collaboration during our data activism journey. We end the discussion by noting that for truly promoting societally beneficial data arrangements, work is needed to circumvent the individualistic and data-centric biases of initiatives such as the MyData.

While we describe the ideological and political underpinnings of data activism, paying attention to social expectations and imaginaries, the activist roles and positions that social scientists can take become clearer. We suggest that in order to make their stance understandable in data activism circles, social scientists need to be aware of the strengths and limitations of their social imaginaries

in order to engage in cross-professional dialogue. Social scientists should also refine their critical faculties, for instance when addressing questions concerning citizenship, participation, dignity, inequality, and discrimination. Data activists generally expect empirically grounded and easily communicated suggestions of how harmful developments could be identified and overcome; as Sarah Pink and Vaike Fors (2017) have suggested, in order for digital data to become a part of processes of change, data practices need to be aligned with 'the generative processes of everyday life'. With a focus on ordinary people, professionals, and communities of practice, ethnographic methods and practice-based analysis can deepen understandings of datafication by revealing how data and associated technologies are taken up, valued, enacted, and sometimes repurposed in ways that either do not comply with imposed data regimes, or that mobilize data in inventive ways (Nafus and Sherman, 2014).

By learning about everyday data work and actual material data practices, humanities and social science scholars can strengthen the understanding of how data technologies could become a part of promoting and enacting more responsible data futures. Paradoxically, in order to arrive at an understanding of how data initiatives support societally beneficial developments, we argue that *non-data-centric* data activism is called for. By aiming at non-data-centric data activism, we can continue to argue against technological solutionism in ways that are critical, but do not deny the possible value of digital data in future-making. The non-data-centric data activism meshes critical thinking into the mundane realities of everyday practices and calls for historically informed and collectively oriented alternatives and action. We suggest that non-data-centric data activism is a form of data activism that can act imaginatively with and within data initiatives to develop new concepts, frameworks and collaborations in order to better steer them.

References

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data & Society*, Oct.

Belli, L., Schwartz, M., & Louzada, L. (2017). Selling your soul while negotiating the conditions: from notice and consent to data control by design. *Health and Technology*, 1-15.

Børsen, T. & Botin, L. (eds) (2013). *What Is Techno-Anthropology?* Aalborg, Denmark: Aalborg University Press.

Kish, L. J., & Topol, E. J. (2015). Unpatients: why patients should own their medical data. *Nature biotechnology*, 33(9), 921-924.

Mayer-Schönberger, V., and K. Cukier. (2013). *Big data: a revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

McQuillan, D. (2016). Algorithmic Paranoia and the Convivial Alternative. *Big Data and Society* 3(2).

McStay, Andrew (2013). *Privacy and Philosophy: New Media and Affective Protocol*. New York: Peter Lang.

Milan, S., & Velden, L. V. D. (2016). The alternative epistemologies of data activism. *Digital Culture & Society*, 2(2), 57-74.

Nafus, D. and Sherman, J. (2014). This One Does Not Go Up to 11: The Quantified Self Movement as an Alternative Big Data Practice. *International Journal of Communication* 8: 1784-1794.

Poikola, A.; Kuikkaniemi, K.; & Kuittinen, O. (2014). My Data – Johdatus ihmiskeskeiseen henkilötiedon hyödyntämiseen [‘My Data – Introduction to Human-centred Utilisation of Personal Data’]. Helsinki: Finnish Ministry of Transport and Communications.

Poikola, A.; Kuikkaniemi, K.; & Honko, H. (2015). MyData – a Nordic Model for Human-centered Personal Data Management and Processing. Helsinki: Finnish Ministry of Transport and Communications.

Raley, R. (2013). Dataveillance and Counterveillance, in ed. Gitelman, *Raw Data is an Oxymoron*. Cambridge: MIT Press.

Ruckenstein, M. & Pantzar, M. (2015). Datafied life: Techno-anthropology as a site for exploration and experimentation. *Techné: Research in Philosophy & Technology* 19(2), 191–210.

Ruckenstein, M., & Schüll, ND (2017). The Datafication of Health. *Annual Review of Anthropology*, *46*, 261-278.

Sharon, T. (2016) Self-Tracking for Health and the Quantified Self: Re-Articulating Autonomy, Solidarity, and Authenticity in an Age of Personalized Healthcare. *Philosophy & Technology*, 1-29.

Taylor, C. (2004). *Modern Social Imaginaries*. Duke University Press.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance and Society* 12(2): 197–208

Van Dijck, J., & Poell, T. (2016). Understanding the promises and premises of online health platforms. *Big Data & Society*, 3(1), 1-11.

Wilbanks, J. T., & Topol, E. J. (2016). Stop the privatization of health data. Nature, 535, 345-348.

Winner, L. (1978). *Autonomous Technology – Technics-out-of-Control As a Theme in Political Thought*. Cambridge, Massachusetts, & London: The MIT Press.

Zuboff, S. (2015). "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30: 75–89.

Wikidocumentaries

Short paper at DHN2018

Susanna Ånäs

background

Wikidocumentaries is a concept for a collaborative online space for gathering, researching and remediating cultural heritage items from memory institutions, open platforms and the users. The setup brings together communities of interest and of expertise to work together on shared topics with open online tools.

The platform will make available collections of open content – images, written documents, maps, data, sounds, print – through querying open repositories, such as the national content aggregator *finna.fi*, *Wikimedia Commons or Flickr Commons*. Wikidocumentaries uses open structured data from *Wikidata* as a means of organizing and filtering the content. Open tools on the platform are used to identify, arrange and enrich the content, but also to remix and reuse it creatively. The resulting data is recirculated to open platforms and made universally available.

For the memory organization, Wikidocumentaries offers a platform for crowdsourcing, for the amateur and expert researchers it provides a community of peers and audiences as well as tools for discovery and interpretation, and from the point of view of the open environments, it acts as an intermediate phase of curation.

Current environments fall short in serving this purpose. Content aggregators focus on gathering, harmonizing and serving the content metadata. Commercial services fail to take into account the open and connected environment in the search for profit. Research environments do not prioritize public access and broad participation. Many participatory projects live short lives from enthusiastic engagement to oblivion due to lack of planning for the sustainability of the results. Wikidocumentaries will try to battle these challenges.

The project has been prepared with a network of participants from Finnish memory institutions, universities and NGOs, who focus on different aspects of the project. Some provide content, while others advance the platform through research or as partners for community projects. This year Wikidocumentaries is supported by the Kone Foundation for co-designing and developing the platform and launching the first community projects.

This short paper will create an initial inventory of research topics that this environment surfaces.

A public domain caleidoscope

Wikidocumentaries launches with Finnish places, people and events that are stored in Wikidata. Each topic is displayed as a page that gathers all open content about the topic from a variety of services. New topics can be created and they will be stored in Wikidata whenever possible.

Technologically Wikidocumentaries uses linked open data and especially Wikidata in structuring cultural heritage content. It creates a proof a concept of creating a federated Wikibase instance, i.e. a Wikidata-like repository that communicates with Wikidata proper. Best practises will be researched in collaboration with other Wikibase projects, such as the Rhizome.org or the Finnish Names Archive project (Kotus), which has been created by members of the project team.

Public & personal

For Wikidocumentaries everyone is notable and differing views are allowed. Each participant can focus on their own viewpoint and research, adding their own pictures and testimonies to topics of their interest.

As the 20th century in Europe is already a “black hole” in the online world for copyright reasons (e.g. Boyle, 2009) a new restricting challenge in bringing cultural heritage online arises with the forthcoming General Data Protection Regulation (European Union). When combining data and images about people in Wikidocumentaries, it must be legally and ethically appropriate. Publicly funded memory organizations’ mission to store and display information about historical subjects will most probably be secured as public or legitimate interest (Voutilainen, 2017), but the situation may be trickier for an amateur online project aggregating cultural heritage content, as it may not fit any specified exception. The museum sector is calling out for the “right to be remembered” as opposed to the right to be forgotten as a means for minorities to be represented in collections and, consequently, in recorded history (Levä, 2017).

In the project we will participate in discussions about technologies of consent, conducted in the MyData community (Conway, 2017). This project may take initiative in innovating novel ways of obtaining permission from living people to display images depicting them.

Other legal challenges can also arise from the European copyright reform. Especially, the proposal to impose neighbouring rights to publishers may complicate aggregating and displaying content from different sources (Communia, 2018).

Content permaculture

New images are saved into Wikimedia Commons and data is stored in Wikidata. Images can also be saved in other partnering projects when saving to Wikimedia projects can not be considered.

New opportunities arise with the advent of structured data for Wikimedia Commons. The world’s largest repository of open content will be remodelled to utilize another federated Wikibase instance, turning all metadata into a structured and machine-readable format. This

will enable systems to take advantage of and contribute to the repository in unprecedented effeciency (Morgan, Fauconnier, 2017).

More generally, Wikidocumentaries will continually revise how it sits as a part of the open de-centralized web. It keeps the tools, the content and the storage independent of one another to allow content to flow freely and prevent vendor lock-in.

Meaningful participation

Wikidocumentaries hosts open source tools to work with the materials. Participants identify and classify, locate, date and transcribe, put the pieces in context and order. Work is done reciprocally, by giving and getting. GLAMs get wide participation in their content, and the content gets curated before being saved in open platforms.

There is a growing body of research on crowdsourcing initiatives within Arts and Humanities projects (e.g. Ridge 2015, Dunn 2012). Research has identified modes of engagement from task-based workflows of crowdsourcing projects (Sprinks et al. 2017) to more open Commons-based peer production of Wikimedia projects (Benkler 2006).

This project aims to position itself as an inclusive environment that is driven by the desire of the participants to collaboratively produce information and make it available in the Cultural Commons. For the participants, it presents a scaffolding approach to participation (Ridge 2015, 160), facilitating moving between microcontributions and a completely open plan.

Wikidocumentaries explores ways to reach out to to underrepresented topics and communities by investigating models of live participation, or “public engagement”, as Susan Schreibman prefers to call participation in the Letters of 1916 project (Schreibman et al. 2017).

Discovery, creativity, play

Wikidocumentaries can tell stories dynamically, for example in automated timelines or Twitter feeds, or you can use the material to remix your own stories. As tools are added, you can create story maps, narrated slideshows or perhaps collectively performed songs. Anyone can add new tools.

Ideally, the platform would grow to become a trusted space to bring together any Open Source tools for enriching distributed cultural heritage assets. Developers can create apps that reuse data produced in other contexts and tools. With this goal in mind, we are inviting the community of stakeholders: content providers, developers, domain experts and more, to play and innovate together in workshops, field trips and hackathons.

Cultural Commons

The work is done together. Amateur researchers, local communities, birds-of-a-feather, relatives, friends and families come together with expert members: museums, libraries, archives and scholars.

Both scholarly and amateur reseachers share similar interests in knowledge discovery and can benefit from work and expertise of one another. Strikingly, the proportion of amateur

researcher visitors in the National Library newspapers service is over 80% (Hölttä 2016), and the results from other platforms also reveal the general interest in historical material. Thus, as a key strategy for further development, we aim to bring amateur and scholarly researchers together to discover historical content and to share and contribute to the Cultural Commons together (e.g. Madison 2009).

References

"Wikidata:Main Page – Wikidata." https://www.wikidata.org/wiki/Wikidata:Main_Page.

"Wikidocumentaries." https://wikidocumentaries.github.io/.

Benkler, Yochai, and Helen Nissenbaum. "Commons-Based Peer Production and Virtue." *Journal of Political Philosophy* 14, no. 4 (December 2006): 394–419. https://doi.org/10.1111/j.1467-9760.2006.00235.x.

Boyle, James. "Google Books and the Escape from the Black Hole | The Public Domain |," September 6, 2009. http://www.thepublicdomain.org/2009/09/06/google-books-and-the-escape-from-the-black-hole/.

Communia Association. "Seven Ways to Save the EU Copyright Reform Effort in 2018 - International Communia Association." Accessed March 7, 2018. https://www.communia-association.org/2018/01/08/seven-ways-save-eu-copyright-reform-effort-2018/.

Conway, Shaun. "How to Operationalise Consent – MyData – Medium." Accessed March 6, 2018. https://medium.com/mydata/how-to-operationalise-consent-7d6d6357d52b.

Dunn, Stuart, and Mark Hedges. "Crowd-Sourcing Scoping Study : Engaging the Crowd with Humanities Research." Centre for e-Research, Department of Digital Humanities, King's College London, 2012. http://crowds.cerch.kcl.ac.uk/wp-content/uploads/2012/12/Crowdsourcing-connected-communities.pdf.

European Union. "Home Page of EU GDPR." https://www.eugdpr.org/.

Hölttä, Tiina. "Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat," 2016. http://tampub.uta.fi/handle/10024/98714.

Kotus. "Nimiarkisto." https://nimiarkisto.fi/.

Levä, Kimmo. "Suomen Museoliiton P.S.-Blogi: Oikeus Tulla Muistetuksi," March 2016. http://museoliitto.blogspot.fi/2016/03/oikeus-tulla-muistetuksi.html.

Madison, Michael J., Brett M. Frischmann, and Katherine J. Strandburg. "Constructing Commons in the Cultural Environment." *Cornell L. Rev.* 95 (2009): 657.

Morgan, Jonathan, and Sandra Fauconnier. “What Galleries, Libraries, Archives, and Museums Can Teach Us about Multimedia Metadata on Wikimedia Commons – Wikimedia Blog,” January 29, 2018. https://blog.wikimedia.org/2018/01/29/glam-multimedia-metadata-commons/.

Morgan, Jonathan. “Research:Supporting Commons Contribution by GLAM Institutions – Meta,” 2017. https://meta.wikimedia.org/wiki/Research:Supporting_Commons_contribution_by_GLAM_institutions.

Open Knowledge Foundation. “OpenGLAM Principles | OpenGLAM.” Accessed February 7, 2018. https://openglam.org/principles/.

Ridge, Mia, ed. *Crowdsourcing Our Cultural Heritage*. Digital Research in the Arts and Humanities. Farnham, Surrey, England: Ashgate, 2014.

———— “Making Digital History. The Impact of Digitality on Public Participation and Scholarly Practices in Historical Research.” The Open University, 2015.

Schreibman, Susan, Vinayak Das Gupta, and Neale Rooney. “Notes from the Transcription Desk: Visualising Public Engagement.” *English Studies* 98, no. 5 (July 4, 2017): 506–25. https://doi.org/10.1080/0013838X.2017.1333754.

Sprinks, James, Jessica Wardlaw, Robert Houghton, Steven Bamford, and Jeremy Morley. “Task Workflow Design and Its Impact on Performance and Volunteers’ Subjective Preference in Virtual Citizen Science.” *International Journal of Human-Computer Studies* 104 (August 1, 2017): 50–63. https://doi.org/10.1016/j.ijhcs.2017.03.003.

Voutilainen, Tomi. “EU:n uusi tietosuoja-asetus – Mikä oikeasti muuttuu? : Säilyttää vai arkistoida – Siinä kysymys.” presented at Digitalia Summer School 2017.

Abstract
Digital Humanities in the Nordic Countries 2018

Medieval Publishing from c. 1000 to 1500
Poster

Medieval Publishing from c. 1000 to 1500 (MedPub) is a five-year project funded by the European Research Council, based at Helsinki University, and running from 2017 2022. The project seeks to define the medieval act of publishing, focusing on Latin authors active during the period from c. 1000 to 1500. MedPub's research hypothesis is that publication strategies were not a constant but were liable to change, and that different social, literary, institutional, and technical milieux fostered different approaches to publishing. A part of the project is to establish a database, whose working title is Medieval Publication Database. The poster will present the main aspects of the projected database and the process of data-gathering.

For the purposes of this research, we define 'publishing' as a social act, involving at least two parties, an author and an audience, not necessarily always brought together. The former prepares a literary work and then makes it available to the latter. Medieval publishing was probably more often a more complex process. It could engage more parties than the two, such as commentators, dedicatees, and commissioners. The social status of these networks ranged from mediocre to grand. They could consist of otherwise unknown monks; or they could include popes and emperors.

We propose that the composition of such literary networks was broadly reactive to large-scale societal and cultural changes. If so, networks of publishing can serve as a vantage point for the observation of continuity and change in medieval societies. We believe such a proposition is significant for the reason that statistical investigations into large-scale social phenomena in the Middle Ages are very rare in the absence of relevant evidence. Therefore, we shall collect and analyse an abundance of data of publishing networks in order to trace how their composition in various contexts may reflect the wider world. It is that last-mentioned aspect that is the key concern of Medieval Publication Database.

It is a central fact for this undertaking that medieval works very often include information on dedication, commission, and commendation; and that, more often than not, this evidence is uncomplicated to collect because the statements in question tend to be short and uniform and they normally appear in the prefaces and dedicatory letters with which medieval authors often opened their works. What is more, such accounts

manifestly indicate a bond between two or more parties. As a rule these parties are the author and a dedicatee and/or a commissioner.

The evidence in question can be collected in the quantities needed for large-scale analysis. The evidence can also be processed electronically and approached statistically. For the function and form of medieval references to dedication and commission remained largely a constant. Eleventh-century dedications resemble those from, say, the fourteenth century. By virtue of such uniformity the data of dedications and commissions may well constitute a unique pool of evidence of social interaction in the Middle Ages. For the data of dedications and commissions can be employed as statistical evidence in *various* regional, chronological, social, and institutional contexts, something that is very rare in medieval studies.

The base framework of the database is a roster of medieval Latin authors and their works. The project's first step is to establish that roster by way of text mining; the materials to be mined come from previous printed and electronical catalogues of medieval Latin writers and/or medieval Latin works. Text-mined entries are standardized and converted into a relational database of individual works and authors, with relevant metadata. Each work will be given a unique identifier and will be specified with the information of its genre and its date of composition (as derived from the active years of its author). Additional information about e.g. editions, surviving copies, etc. may be included in case the catalogues to be mined provide us with such. The first catalogue we deal with is *A Handlist of Latin Writers of Great Britain and Ireland before 1540* by R. Sharpe. Its selection as a starting point was on account of two virtues: the catalogue attends to a coherent geographical whole and does so with remarkable consistency.

At the second step, the standardized entries of published works will be enriched with information related to publishing networks. In most cases the information in question concerns dedication and/or commission. To put shortly, we will mainly identify authors, dedicatees, and commissioners. The database records their name, social status and the geographical region and the time frame of their activity. As noted above, such data can relatively easily be gathered from works that have been edited in modern times. At subsequent steps, we will also study manuscripts for works that remain unpublished in print.

Yet there are complications that we must overcome at this initial stage of the project. These mainly pertain to the ontologies and hierarchies needed when defining the social status and geographical region of authors and other persons. In particular the latter, the

geographical aspect of publishing, appears problematic as to standardization. The poster will explicate this aspect in detail.

To conclude, our metadata to be harvested and enriched can be summarised as temporal, spatial, social, and literary. We believe that once standardized such data can be used as evidence in statistical inquiries in various fields, ranging from literary studies to social history. At the first stage, the Medieval Publication Database will cover medieval Latin authors connected to the British Isles. The final goal is to encompass medieval Latin Europe in its entirety. The database will be published online in accordance with the Open Science guidelines.

The poster will introduce the database, metadata scheme and how the data of dedications and commissions will be harvested.

Daria Glebova

PhD Candidate
Institute of Slavic Studies, Russian Academy of Sciences

Using *rolling.classify* on the Sagas of Icelanders: Collaborative Authorship in *Bjarnar saga Hítdælakappa*

This poster will present the results of an application of the *rolling.classify* function in *Stylo (R)* to a source of unknown authorship and extremely poor textual history – *Bjarnar saga Hítdælakappa,* one of the "family sagas" or "sagas of Icelanders" (Ice. *Íslendingasögur*). This case study sets the usual *Stylo* authorship attribution goal aside and concentrates on the composition of the main vitness of *Bjarnar saga*, ms. AM 551 d α, 4to (17th c.), which was the source for the most existing *Bjarnar saga* copies. It aims not only to find and visualise new arguments for the working hypothesis about the AM 551 d α, 4to composition but also to touch upon the main questions that rise before a student of philology daring to use *Stylo* on the Old Icelandic saga ground, i.e. what *Stylo* tells us, what it does not, and how can one use it while exploring the history of a text that exists only in one source.

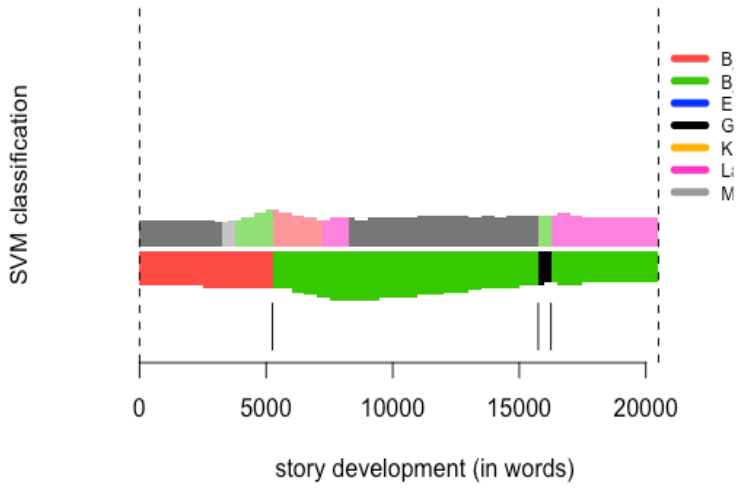
It has been noted that *Bjarnar saga* shows signs of a stylistic change between the first 10 chapters and the rest of the saga – the characters suddenly change their behaviour (Sígurður Nordal 1938, lxxix; Andersson 1967, 137-140), the narrative becomes less coherent and, as it seems, acquires a new logic of construction (Finlay 1990-1993, 165-171). More detailed narrative analysis of the saga showed that there is a difference in the usage of some narrative techniques in the first and the second parts, i.e., for example, the narrator's work with point of view and the amount of their intervention in the saga text (Glebova 2017, 45-57). Thus, the question is – what is the relationship between the first 10 chapters and the rest of *Bjarnar saga*? Is the change entirely compositional and motivated by the narrative strategy of the medieval compiler or it is actually a result of a compilation of two texts that have two different authors?

As it often happens with Old Icelandic sagas in general, the problem aggravates due to the *Bjarnar saga's* poor preservation. There is not much to compare and work with; the most of the saga vitnesses are copies from one 17th c. manuscript, AM 551 d α, 4to (Boer 1893, xii-xiv; Sígurður Nordal 1938, xcv-xcvii; Simon 1966 (I), 19-149). This manuscript also has its flaws: it has two lacunae, one in the very beginning of the saga (ch. 1-5,5 in *ÍF III*) and another in the middle (between ch. 14-15 in *ÍF III*). The second lacuna is irreconstructable while the first one is usually substituted by a fragment from the saga's short reduction that was preserved in copies of 15th c. kings' saga compilation, *Separate saga St. Olaf* in *Bœjarbók* (Finlay 2000, xlvi), and that actually ends right on the 10th chapter of the longer version. It seems that the text of the shorter version is a variant of the longer one (Glebova 2017, 13-17); precise relationships between the short and long redactions, however, are irreconstructable due to the lacuna in AM 551 d α, 4to. The existance of the short version with these particular length and contents is indeed very important to the study of *Bjarnar saga* composition in AM 551 d α, 4to as it creates a chance that the first 10 chapters of AM 551 d α, 4to could exist separately at some point of the *Bjarnar saga*'s text history or at least that these chapters were seen by the medieval compilers as something solid and complete. This is as far as traditional philology can go in this case – the state of the sources does not allow saying more. However, is there anything else that could shed some light on whether these chapters existed separately or were written by the same hand?

In this study it was decided to try sequential stylometric analysis available in *Stylo* package for *R* (Eder, Kestemont, Rybicki 2013) as a function *rolling.classify* (Eder 2015). As we are interested in the different parts of the same text, rolling stylometry seems to be a more preferable method to cluster analysis, which takes the whole text as an entity and compares it to the reference corpus; alternatively, in case with rolling stylometry the text is divided into smaller segments which allows for a deeper investigation of the stylistic variation in the text itself

(Rybicki, Eder, Hoover 2016, 126). To do the analysis a corpus was made from the two parts of *Bjarnar saga* and several other Old Icelandic sagas (not only Sagas of Icelanders, but also Kings' sagas) used as a context; the whole corpus was taken from sagadb.org in Modern Icelandic normalised orthography. Then, a reference set was built out of a sample from the 1st part of *Bjarnar saga* and a sample from the 2nd part as well as samples from other sagas. This reference set was used to train a classification model using support vector machines (SVM). Finally, the model was applied to the whole *Bjarnar saga* that was then tested against the reference set. A series of tests was conducted with different slice sizes (5000 words to 2000) and different amounts of MFW (most frequent words). The preliminary results show that there is a stylistic division in the saga as the style of the first part is not present in the second one at all. See an example in Fig.1 where the red colour corresponds to the 1st part.

Fig.1. Slice.size = 5000, slice.overlap = 4500, MFW = 100



This would be an additional argument for the idea that the first 10 chapters existed separately and were added by the *Bjarnar saga* compiler during the saga construction. *Stylo* counts the most frequent words, which are not so generically specific (like *og*, *að*, etc.); thus, the collaborative authorship still could have taken place. However, the previous results of the narratological analysis make us think further – if the structure of the saga shows traces of very careful planning and mirror composition (Glebova 2017, 18-33), most probably it is not the case of a clumsy compilation. Even if the first part and the second part could exist separately, they were *chosen* to be put in this order and this particular form. So what could the stylistic change mean then? One could argue that it could be a generic division as the first part is set in Norway and deals a lot with St. Olaf; the change of genre could result in the change of style. On the other hand, the reason also could be pragmatic – the first part could be seen as an introduction and the second part as the main narrative. Whatever be the case, while sewing together the existing material the medieval compiler made an effort to create a solid text and this effort is worth studying with more attention.

Bibliography:

Andersson, Theodor M. (1967). The Icelandic Family Saga: An Analitic Reading. Cambridge, MA.

Boer, Richard C. (1893). *Bjarnar saga Hítdælakappa*, Halle.

Eder, M. (2015). "Rolling Stylometry." *Digital Scholarship in the Humanities*, Vol. 31-3: 457–469.

Eder, M., Kestemont, M., Rybicki, J. (2013). “Stylometry with R: A Suite of Tools.” *Digital Humanities 2013: Conference Abstracts*. University of Nebraska–Lincoln: 487–489.

Finlay, A. (1990-1993). “Nið, Adultry and Feud in Bjarnar saga Hítdælakappa.” *Saga-Book of the Viking Society* 23 (1990-1993): 158-178.

Finlay, A. (2000). *The Saga of Bjorn, Champion of the Men of Hitardale*, Enfield Lock.

Glebova D. (2017). *A Case of An Odd Saga. Structure in Bjarnar saga Hítdælakappa.* MA thesis, University of Iceland. Reykjavík (http://hdl.handle.net/1946/27130).

Rybicki, J., Eder, M., Hoover, David L. (2016). “Computational Stylistics and Text Analysis.” In *Doing Digital Humanities: Practice, Training, Research*, edited by Constance Compton, Richard J. Lane, Ray Siemens. London, New York: 123-144.

Sigurður Nordal, and Guðni Jónsson (eds.) (1938). “Bjarnar saga Hítdælakappa.” In *Borgfirðinga sögur*, Íslenzk fornrit 3, 111-211. Reykjavík.

Simon, John LeC. (1966). *A Critical Edition of Bjarnar saga Hítdælakappa*. Vol. 1-2. Unpublished PhD thesis, University of London.

----------POSTER----------

Topics: History

Key words: rolling stylometry, Íslendingasögur, medieval sources, collaborative authorship, composition

Two cases of meaning change in Finnish newspapers, 1820-1910

Antti Kanner

Abstract for Nordic Digital Humanities Conference in Helsinki, March 2018

In Finland the 19th century saw the formation of number of state institutions that came to define the political life of the Grand Duchy and of the subsequent independent republic. Alongside legal, political, economic, and social institutions, Modern Finnish, as an institutionally standardised language, can be seen in this context. As the majority of residents of Finland were native speakers of Finnish dialects, adopting Finnish was necessary for state's purposes in extending its influence within the borders of the autonomous Grand Duchy. Widening domains of use of Finnish also played an important role in the development of Finnish national identity. In the last quarter of 19th century, Finnish started to gain ground as the language of administrative, legal, and political discourses alongside Swedish. It is during this period that we find most of the crucial conceptual processes that shaped Finnish political history.

In this paper I present two related case studies from my doctoral research, where I seek to understand the semantic similarity scores of so-called Semantic Vector Spaces in terms of linguistic semantics. The vector spaces have been obtained from large historical corpora of Finnish newspapers. Historical corpora are best understood as collections of past speech acts and the view they provide to changing meanings of words is shaped by contextual and pragmatic factors present at the moment of a texts' production. For this reason, understanding and explicating the historical context of observed processes is essential when studying temporal dynamics in semantic changes. To this end, I will try to reflect the theoretical side of my work in the light provided by actual cases of historical meaning changes. My research falls under the heading of Finnish Language, but is closely related to intellectual and conceptual history and computational linguistics.

The main data for my research comes from the National Library of Finland's Newspaper Collection, which I use via the KORP service API provided by Language Bank of Finland. The collection contains nearly all newspapers and periodicals published in Finland from 1771 to 1910, and Finnish publications from 1820. The collection is, however, very heterogenous, as the press and other forms of printed public discourse in Finnish only developed during the 19th century. Historical variation in conventions of typesetting, editing, and orthography, as well as paper quality used for printing make it difficult for OCR systems to recognize characters with 100 percent accuracy. Kettunen et. al. estimated that OCR accuracy is actually somewhere between 60 and 80 percent. However, not all problems in the automatic recognition of the data come from OCR problems or historical spelling variation. Much is also due to linguistic factors: the 19th century saw large scale dialectal, orthographical, and lexical variation in written Finnish. To exemplify the scale of variation, when a morphological analyser for Modern Finnish (OMORFI, Pirinen 2015) was used, it could only

parse around 60 percent of the wordlist of the Corpus of Early Modern Finnish (CEMF). For these reasons (unreliable results from automated parser and the temporal heterogeneity inherent in the data), conducting a methodology robust study poses a challenge. To mitigate these issues, the approach chosen was to use a number of analyses and see whether results could be combined to produce a coherent view of the historical change in word use.

All of the analyses in this work are based on term-feature matrices, of which term-document, term-collocation and term-morphological case category can be said to be of different types. Depending on specific tasks, methods in computational linguistics, such as LDA (Blei, Ng & Young 2004) or word2vec (Mikolov et. al. 2013), select one type of term-feature matrix as the starting point and then process this matrix into a more concentrated, embedded vector space. In research, most attention is usually reserved to the algorithm and selection of the type of original feature matrix in usually treated as more or less trivial matter. However, eg. Levy & Goldberg (2014) have noted that linguistic regularities observed in the embedded vector spaces are not consequences of the embedding process, but that the embedding process preserves those regularities well. Also, earlier research claims that the different ways the word representations are built from corpus data (ie. the specific type of the word-feature matrix used) seems to measure different types of semantic relatedness as semantic similarity (Sahlgren 2006, 2008). The methodological choices of this work, then, stem from these two observations. No embedding algorithms are used, and an array of analyses based on different types of word-feature matrices, is composed to monitor different semantic relations for semantic changes.

While a number of analyses based on different types of word-feature matrices were conducted (such as varying ngrams and skip-grams), two analyses deserve further discussion here. First, an analysis based on term-document matrices, such as topic modelling, seems to describe meaning relations that could be said to be associative, schematic, or discursive. In this study, this analysis was conducted using second order collocations (Bertels & Speelman 2014 and Heylen, Wielfaerts, Speelman, Geeraerts 2014) instead of algorithms like LDA, that are widely used for purpose of topic modelling. Preliminary tests showed that in this specific case, LDA was not able to produce well-formed topics in comparison to results from clustering second order collocations. This simpler approach seemed to be robust for some properties in the data that yielded LDA unusable, while it was left unclear what those properties actually were. Second, an analysis based on syntactic features was conducted by substituting syntactic dependencies with case marking distributions. This can be done on the grounds that the case selection in Finnish is mostly governed by syntax, as case selection is used to express syntactic relations between, for example, constituents of nominal phrases or predicate verb and its arguments (Vilkuna 1989). As automated syntactic parsing relies on preprocessed morphological analysis, fragility of syntactic parsing to errors in morphological preprocessing is of second order of magnitude. The aggregated morphological distributions on the other hand seem to be quite robust with regard to mistakes in the data, as the nature of noise the errors introduced is, in most cases, quite uniform. When the task is to track signals

of change, morphological case distributions can be used as sufficient proxies for dependency distributions.

The first of my case studies focuses on the Finnish word *maaseutu*. After its introduction to Finnish in the 1830's, *maaseutu* was used in more variety of related meanings, mostly referring to specific rural areas or communities. Staring from the 1870s it developed into a collective singular, referring to countryside as an undivisive whole and frequently contrasted to the urban, often lexicalised as *kaupunki*, the city. At the time when the the collective singular emerges, we find a number of occurrences which are vague in respect to specificity and collectivity.

Combining information from my analysis to newspaper metadata yields an image of a dynamic situation. The emergence of the collective singular stands out clearly, and is connected to an accompanying discourse of negotiating urban-rural relations on a national instead of regional level. This change can be pinpointed quite precisely to 1870s, and to newspapers with geographically wider circulation and a more national identity.

The second word of interest is *vaivainen*, an adjective referring to a person or a thing either being of wretched or inadequate quality, or suffering from a physical or mental ailment. When used as a noun, it refers to a person of very low and excluded social status and extreme poverty. The word has a biblical background, being used in older Finnish Bible translations (in, for example, the Sermon on the Mount as the equivalent of *poor* in Matt. 5:13: "blessed are the poor in spirit"), and as such was a natural choice to name the recipients of church charities. When the state poverty relief system started to take its form in the mid 19th century, it was built on top of earlier church organizations (Von Aerschot 1996), thus church terminology was carried over to these state institutions. Today, however, the word appears in Modern Finnish mostly in poetically archaic or historical contexts. It has disappeared from the vocabulary of social policy or social legislation by the early 20th century.

When tracking the contexts of the word over the 19th century using context word clusters based on second order collocations, two clear discoursal trends appear: the poverty relief discourse, that already in the 1860's is pronounced in the data, disperses into a complex network of different topics and discursive patterns. As state run poverty relief institutions become more complex and efficiently administered, the moral foundings of the whole enterprise are discussed alongside reports of everyday comings and goings of individual institutions or, indeed, tales of individual relief recipient's fortunes. The other trend involves the presence of religious or spiritual discourse which, against preliminary assumptions does not wane into the background, but experiences a strong surge in the 1870s and 1880s. This can be explained in part by the growth of revivalist Christian publications in the National Library Corpus, but also by the intrusion of Christian connotations into the political discussion on poverty relief systems. It is as if the word *vaivainen* functions as a kind of lightning rod of Christian morality in public poverty relief discourse.

While the methodological contributions of this paper are not highly ambitious in terms of language technology or computational algorithms, the combination of a complementary array of analyses instead of a methodology based on a single highly complex and opaque algorithm, might be seen to show an innovative approach to Digital Humanities. I argue that robustness and simplicity of methods makes the overall workflow more transparent, and this transparency makes it easier to interpret the results in wider historical or linguistic contexts. This allows us to ask questions which are not confined to the fields of computational linguistics or lexical semantics, but apply to wider areas of Humanities scholarship. This shared relevance of questions, intersections of interests of knowledge, to my understanding, lies at the core of Digital Humanities.

References

Bertels, A. & Speelman, D. (2014). "Clustering for semantic purposes. Exploration of semantic similarity in a technical corpus." *Terminology* 20:2, pp. 279–303. John Benjamins Publishing Company.

Blei, D., Ng, A. Y. & Jordan, M. I. (2003). "Latent Dirichlecht Allocation." *Journal of Machine Learning Research* 3 (4–5). Pp. 993–1022.

CEMF, *Corpus of Early Modern Finnish.* Centre for Languages in Finland. http://kaino.kotus.fi

Heylen, C., Peirsman Y., Geeraerts, D. & Speelman, D. (2008). "Modelling Word Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms." *Proceedings of LREC 2008*.

Huhtala, H. (1971). *Suomen varhaispietistien ja rukoilevaisten sanankäytöstä : semanttis-aatehistoriallinen tutkimus*. [On the vocabulary of the early Finnish pietist and revivalist movements]. Suomen Teologinen Kirjallisuusseura.

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T. & Kervinen, J. (2014). "Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods". In *IFLA World Library and Information Congress Proceedings : 80th IFLA General Conference and Assembly*. Lyon. France.

Levy, O. & Goldberg, Y. (2014): "Linguistic Regularities in Sparse and Explicit Word Representations." In *Proceedings of the Eighteenth Conference on Computational Language Learning.* Pp. 171-180.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. 2013: "Efficient Estimation of Word Representations in Vector Space." In arXiv preprint arXiv:1301.3781.

Pirinen, T. (2015). "Omorfi—Free and open source morphological lexical database for Finnish". In *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015*.

Sahlgren, M. (2006): *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* SICS. Stockholm University.

Sahlgren, M. (2008): "Distributional Semantics". *Rivista di Linguistica* 20:1. Pp. 33-53.

Vilkuna, M. (1989). *Free word order in Finnish: Its syntax and discourse functions*. Suomalaisen Kirjallisuuden Seura.

Von Aerschot, P. (1996). *Köyhät ja laki: toimeentukilainsäädännön kehittyminen kehitys oikeudellistusmisprosessien valossa.* [The poor and the law: development of Finnish welfare legislation in light juridification processes.] Suomalainen Lakimiesyhdistys.

3rd DHN Conference
7–9 March 2018, Helsinki

Maria Kallio
the National Archives of Finland
maria.kallio@arkisto.fi

Handwritten Text Recognition and 19th century Court Records

This paper will demonstrate how the READ project is developing new technologies that will allow computers to automatically process and search handwritten historical documents. These technologies are brought together in the Transkribus platform, which can be downloaded free of charge at https://transkribus.eu/Transkribus/. Transkribus enables scholars with no in-depth technological knowledge to freely access and exploit algorithms which can automatically process handwritten text. Although there is already a rather sound workflow in place, the platform needs human input in order to ensure the quality of the recognition. The technology must be trained by being shown examples of images of documents and their accurate transcriptions. This helps it to understand the patterns which make up characters and words. This training data is used to create a Handwritten Text Recognition model which is specific to a particular collection of documents. The more training data there is, the more accurate the Handwritten Text Recognition can become.

Once a Handwritten Text Recognition model has been created, it can be applied to other pages from the same collection of documents. The machine analyses the image of the handwriting and then produces textual information about the words and their position on the page, providing best guesses and alternative suggestions for each word, with measures of confidence. This process allows Transkribus to provide the automatic transcription and full-text search of a document collection at high levels of accuracy.

For the quality of the text recognition, the amount of training material is paramount. Current tests suggest that models for specific style of handwriting can reach a Character Error Rate of less than 5%. Transcripts with a Character Error Rate of 10% or below can be generally understood by humans and used for adequate keyword searches. A low Character Error Rate also makes it relatively quick and easy for human transcribers to correct the output of the Handwritten Text Recognition engine. These corrections can then be fed back into the model in order to make it more accurate. These levels also compare favorably with Optical Character Recognition, where 95-98% accuracy for early prints is possible.

Of even more interest is the fact that a well-trained model is able to sustain a certain amount of differences in handwriting. Therefore, it can be expected that, with a large amount of training material, it will be possible to recognize the writing of an entire epoch (e.g. eighteenth-century English writing), in addition to that of specific writers.

The case study of this paper is the Finnish court records from the 19th century. The notification records which contain cases concerning guardianships, titles and marriage settlements, form an enormous collection of over 600 000 pages. Although the material is in digital form, the usability is still poor due to the lack of indices or finding aids. The National Archives of Finland started to produce transcripts from a small part of the collection in cooperation with the Genealogical Society of Finland in 2017. Around 30 volunteers have been producing ground truth for training of Handwritten Text Recognition. The first HTR model was trained based on 75,000 words of training data. The result was very good, with an average character error rate (CER) of only 12 per cent. The genealogists have continued to work on the material and the goal is to collect as much training data as possible in order to reach a CER of 5–10 per cent. Even with the current result of CER 12% it is possible to search and use the material with the Key Word Spotting -tool in Transkribus, which already improves the research opportunities of the court records considerably. The goal is to provide at least part of the notification records in computer readable form by the end of the project and thereby enable new versatile possibilities for research.

Creating a corpus of communal court minute books: a challenge for digital humanities

Gerth Jaanimäe, Liina Lindström, Kadri Muischnek, Siim Orasmaa, Maarja-Liisa Pilvik (University of Tartu),
Kersti Lust (The National Archives of Estonia)

This paper presents the work of a digital humanities project concerned with the digitization of Estonian communal court minute books. The local communal courts in Estonia came into being through the peasant laws of the early 19th century and were the first instance class-specific courts, that tried peasants. Rather than being merely judicial institutions, the communal courts were at first institutions for the self-government of peasants, since they also dealt with police and administrative matters. After the municipal reform of 1866, however, the communal courts were emancipated from the noble tutelage and the court became a strictly judicial institution, that tried peasants for their minor offences and solved their civil disputes, claims and family matters. The communal courts in their earlier form ceased to exist in 1918, when Estonia became independent from the Russian rule.

The National Archives of Estonia holds almost 400 archives of communal courts from the pre-independence period. They have been preserved very unevenly and not all of them include minute books. The minute books themselves are also written in an inconsistent manner, the earlier minute books are often written in German and the writing is strongly dependent on the skills and will of the parish clerk. However, the materials from the period starting with the year 1866, when the creation of the minute books became more systematic, are a massive and rich source shedding light on the everyday lives of the peasantry. Still, at the moment, the users of the minute books meet serious difficulties in finding relevant information since there are no indexes and one has to go through all the materials manually. The minute books are also a fascinating resource for linguists, both dialectologists and computational linguists: the books contain regional varieties tied to specific genre and early time period (making it possible to detect linguistic expressions, which are rare in atlases, for example, and also in dialect corpus, which represents language from about 100 years later) while also being a written resource, reflecting the writing traditions of the old spelling system. This is also what makes these texts complex and challenging for automatic analysis methods, which are otherwise quite well-established in contemporary corpus linguistics.

In our talk we present a project dealing with the digitization and analysis of the minute books from the period between 1866 and 1890. The texts were first digitized in the 2000s and preserved in a server in html-format, which is good for viewing, but not so good for automatic processing. After the server crashed, the texts were rescued via web archives and the structure of the minute books was used to convert the documents automatically into a more functional format using xml-markup and separating the body text with tags referring to information about the titles, dates, indexes, participants, content and topical keywords, which indicate the purview of the communal courts in that period.

We discuss the workflow of creating a digital resource in a standardized and maximally functional format as well as challenges, such as automatic text processing for cleaning and annotating the corpus in order to distinguish the relevant layers of information. Tools developed for Estonian morphological analysis are trained on contemporary written standard Estonian. Communal court minute books, however, include language variants, which are a mixture of dialectal language, inconsistent spelling and the old spelling system. In the presentation, we introduce the results of our first attempts to apply the automatic text analysis tools to the materials of communal court minute books, the problems that we've run into, and provide solutions for overcoming these problems. Similar experiments of testing tools developed for contemporary language on historical texts have been conducted on other languages as well, e.g. Icelandic (Lofsson 2013; Rögnvaldsson and Helgadóttir 2011), German (Scheible et al. 2011; Bollmann 2013), Swedish (Petterson 2016), and Spanish (Sánchez-Marco et al. 2011). To achieve better accuracy rate for tagging, the two main proposed solutions have been text normalization (Scheible et al. 2011; Bollmann 2013; Petterson 2016) and tool adaption (Rögnvaldsson and Helgadóttir 2011; Sánchez-Marco et al. 2011).

As the National Archives have a considerable amount of communal court minute books, which are thus far only in a scanned form, the digitized minute books collection is planned to expand using crowdsourcing oportunities. The final aim of the project is to create a multifunctional source, which could be of interest for researchers of different fields within the humanities. In addition to comprising important lingustic information, it enables systematic studying of family matters, taxes, granary loans, old age support, tenancy, damages, contractual relationships, credit relations, inventories, living conditions, and minor criminal offences. Furthermore, the database can be linked to individual level demographic databases as well as genealogical databases of various sorts. Thus, both professional and hobby researchers have a great potential to benefit from the project. For example, identification of the kin relationships of otherwise 'anonymous' people as well as of their household and class affiliation allows analyzing the events recorded in the minute books in terms of class, kin and networks.

Preliminary analyses

In order to enable queries with different degrees of specificity in the corpus, the texts also need to be linguistically analyzed. For historical texts written in languages with small number of inflectional forms per lemma, handling spelling variation and canonicalization (Piotrowski 2012: 73-78) is sufficient for word and string based searches, frequency lists etc. However, for languages with rich morphological system that is not enough and the text needs to be lemmatized, which in case of rich inflectional morphology, is done by performing morphological analysis and disambiguation. For both named entity recognition (NER), which enables network analysis and links the events described in the materials to geospatial information, and morphological annotation, which makes it possible to perform queries based on lemmas or grammatical information, we have applied the EstNLTK library (Orasmaa et al. 2016) in Python, which is developed for processing contemporary written standard Estonian. NER's performance was satisfactory, i.e. it found recognized names well, even though it systematically overrecognized organization names. The most complicated issue so far has been the morphological analysis of word forms. Apart from being simply an older version of

Estonian, the language represented by Estonian communal court minute books contains a lot of twofold variation: spelling variation and dialectal variation. Estonian dialects are divided into two main groups: Northern and Southern; Standard Estonian is based on the Northern dialects. Although the parish clerk not necessarily was a speaker of the local dialect, this seems quite often to have been the case. One may therefore assume that automatic morphological analysis and lemmatization of the texts from the Northern parishes should give better results than that of the texts from the Southern parishes.

In order to roughly estimate the quality of the outcome of morphological analysis of communal court minute books, we processed the texts using EstNLTK morphological analyzer without the guesser. Running the analyser without the guesser means that the out-of-vocabulary words, apart from compounds and regular derivations, are tagged as unknown words. The percentage of unknown words per parish is presented in Table 1.

Table 1. Results of automatic morphological analysis for communal court minute books

Parish	Dialectal group	% of words with morphological analysis	of them unambiguous, %	% of unknown words	of them with capital initial letter, %
Laiuse	Northern	93	49	7	74
Mihkli	Northern	90	50	10	56
Navesti	Northern	89	48	11	38
Uue-Suislepa	Southern	86	51	14	38
Alatskivi	Northern	84	54	16	46
Kiuma	Southern	83	53	17	35
Maasi	Northern	83	46	17	34
Vastse-Nõo	Southern	81	53	19	27
Laeva	Southern	80	53	19	43
Aru	Southern	79	53	21	34
Tarvastu	Southern	75	56	25	23
Pangodi	Southern	69	54	31	27
Kahkva	Southern	67	57	33	26

Kärevere	Southern	61	56	39	29
Mäksa	Southern	61	59	39	21
Joosu	Southern	61	48	39	24
Haaslava	Southern	60	61	40	24
Kokora	Southern	60	51	40	29
Valguta	Southern	59	61	41	16
Luke	Southern	53	64	47	27
Suure-Konguta	Southern	49	57	51	27
Väike-Rõngu	Southern	48	53	52	18

The percentage of recognized words varies from 93% for the texts of Laiuse belonging to the Northern dialect group to 48% for the texts from Väike-Rõngu belonging to the Southern dialect group. This perhaps means that due to wide inter-parish variation, we will have to use different means for lemmatizing and/or normalizing texts from different parishes or parish groups.

For comparison, we also processed fiction and newspaper texts in present-day Standard Estonian. The results are presented in Table 2. The comparison of these results with those of the parish court texts shows that the analyzer recognizes minimally *ca*. 4% more word forms when analysing present-day Estonian texts and out of the out-of-vocabulary words *ca*. 12% more word forms begin with a capital letter, i.e. are probably proper nouns.

Table 2. Results of automatic morphological analysis for present-day Standard Estonian texts

text class	% of words with morphological analysis	of them unambiguous, %	% of unknown words	Of them with capital initial letter, %
fiction	98	54	2	86
newspaper	97	60	3	87

References:

Bollmann, Marcel (2013). POS Tagging for Historical Texts with Sparse Training Data. - Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pp 11–18.

Loftsson, Hrafn (2013). Tagging the Past: Experiments Using the Saga Corpus. - Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA-2013).

Orasmaa, Siim, Timo Petmanson, Alexander Tkachenko, Sven Laur, Heiki-Jaan Kaalep (2016). ESTNLTK - NLP Toolkit for Estonian. - Proceedings of LREC 2016, pp 2460–2466.

Petterson, Eva (2016). Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. Doctoral Thesis, Uppsala University.

Piotrowski, Michael (2012). Natural Language Processing for Historical Texts. Morgan & Claypool Publishers.

Rögnvaldsson, Eirikur and Sigrún Helgadóttir (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. - C. Sporleder, A. van den Bosch, and K. Zervanou, editors, Language Technology for Cultural Heritage, Theory and Applications of Natural Language Processing, pp 63–76.

Sánchez-Marco, Cristina, Gemma Boleda and Lluís Padró (2011). Extending the tool, or how to annotate historical language varieties. - Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pp 1–9.

Scheible, Silke, Richard J. Whitt, Martin Durrell and Paul Bennett (2011). Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. - Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pp 19–23.

Extracting script features from a large corpus of handwritten documents

Before the advent of the printing press, the only way to create a new piece of text was to produce it by hand. The medieval text culture was almost exclusively a handwritten one, even though printing began towards the very end of the Middle Ages. As a consequence of this, the medieval text production is very much characterized by variation of various kinds: regarding language forms, regarding spelling and regarding the shape of the script. In the current presentation, the shape of the script is in focus, an area referred to as palaeography. The introduction of computers has changed this discipline radically, as computers can handle very large amounts of data and furthermore measure features that are difficult to deal with for a human researcher.

In the current presentation, we will demonstrate two investigations within digital palaeography, carried out on the medieval Swedish charter corpus in its entirety, to the extent that this has been digitized. The script in approximately 14 000 charters has been measured and accounted for, regarding aspects described below. The charters are primarily in Latin and Old Swedish, but there are also a few in Middle Low German. The overall purpose for the investigations is to search for script features that may be significant from the perspective of separating one scribe from another, i.e. scribal attribution. As the investigations have been done on the entire available charter corpus, it is possible to visualize how each separate charter relates to all the others, and furthermore to see how the charters may divide themselves into clusters on the basis of similarity regarding the investigated features.

The two investigations both focus on aspects that have been looked upon as significant from the perspective of scribal attribution, but that are very difficult to measure, at least with any degree of precision, without the aid of computers. One of the investigations belongs to a set of methods often referred to as Quill Features. This method focuses, as the name states, on how the scribe has moved the pen over the script surface (parchment or paper). The medieval pen, the quill, consisted of a feather that had been hardened, truncated and split at the top. This construction created variation in width in the strokes constituting the script, mainly depending on the direction in which the pen was moved, and also depending on the angle in which the scribe had held the pen. This is what this method measures: the variation between thick and thin strokes, in relation to the angle of the pen. This method has been used on medieval

Swedish material before, namely a medieval Swedish manuscript (Cod. Ups. C 61, 1104 pages), but the current investigation accounts for ten times the size of the previous investigation, and furthermore, we employ a new type of evaluation (see below) of the results that to our knowledge has not been done before.

The second investigation focuses on the relations between script elements of different height, and the proportions between these. For instance three different formations can be discerned among the vertical scripts elements: minims (e.g. in 'i', 'n' and 'm'), ascenders (e.g. in 'b', 'h' and 'k') and descenders (e.g. in 'p' and 'q'). The ascender can extend to a various degree above the minim, and the descender can extend to a various degree below the minim, creating different proportions between the components. These measures have also been extracted from the entire available medieval Swedish charter corpus, and display very interesting information from the perspective of scribal identity. It should be noted that the first line of a charter often is divergent from the rest of the charter in this respect, as the ascenders here often extends higher than otherwise. In a similar way, the descenders of the last line of the charters often extend further down below the line as compared to the rest of the charter. In order for a representative measure to be gained from a charter, these two lines must be disregarded.

One of the problems when investigating individual scribal habits in medieval documents is that we rarely know for certain who has produced them, which makes the evaluation difficult. In most cases, the scribe of a given document is identified through a process of scribal attribution, usually based on palaeographical and linguistic evidence. In an investigation on individual scribal features, it is not desirable to evaluate the results on the basis of previous attributions. Ideally, the evaluation should be done on charters where the identity of the scribe can be established on external features, where his/her identity is in fact known. For this purpose, we have identified a set of charters where this is actually the case, namely where the scribe himself/herself explicitly states that he/she has held the pen (in our corpus, there are only male scribes). These charters contain a so-called scribal note, containing the formula *ego X scripsi* ('I X wrote'), accompanied by a symbol unique to this specific scribe. One such scribe is Peter Tidikesson, who produced 13 charters with such a scribal note in the period 1432–1452, and another is Peter Svensson, who produced six charters in the period 1433–1453. This selection of charters is the means by which the otherwise big data-focused computer aided methods can be evaluated from a qualitative perspective. This step of

evaluation is crucial in order for the results to become accessible and useful for the users of the information gained.

Seppo Eskola & Lauri Leinonen
National Archives of Finland
Digital humanities in the Nordic Countries 2018
Abstract, long presentation format

Diplomatarium Fennicum and the digital research infrastructures for medieval studies

Digital infrastructures for medieval studies have advanced in strides in Finland over the last few years. Most literary sources concerning medieval Finland − the Diocese of Åbo − are now available online in one form or another: Diplomatarium Fennicum encompasses nearly 7 000 documentary sources, the Codices Fennici project recently digitized over 200 mostly well-preserved pre-17th century codices and placed them online, and Fragmenta Membranea contains digital images of 9 300 manuscript leaves belonging to over 1 500 fragmentary manuscripts. In terms of availability of sources, the preconditions for research have never been better. So, what's next?

This presentation discusses the current state of digital infrastructures for medieval studies and their future possibilities. For the past two and a half years the presenters have been working on the Diplomatarium Fennicum webservice, published in November 2017, and the topic is approached from this background. Digital infrastructures are being developed on many fronts in Finland: several memory institutions are actively engaged (the three above-mentioned webservices are developed and hosted by the National Archives, the Finnish Literature Society, and the National Library respectively) and many universities have active medieval studies programs with an interest in digital humanities. Furthermore, interest in Finnish digital infrastructures is not restricted to Finland as Finnish sources are closely linked to those of other Nordic countries and the Baltic Sea region in general.

The technical solutions used in different medieval studies projects are very varied − both between the Finnish services and internationally. Metadata models can be based on e.g. Dublin Core, xml or tailor-made solutions, and even services with very similar goals have rarely made the same technical choices in general. Good international points of comparison for Diplomatarium Fennicum are for instance Diplomatarium Suecanum (SDHK), Diplomatarium Danicum, and the webservice of the 'Making of Charlemagne's Europe' project, all of which have been examined when making Diplomatarium Fennicum. What perhaps makes the Finnish services unique, and could make them of special interest internationally, is the fact that surviving literary sources concerning medieval Finland are limited in scope to a degree that has made it possible to include them all in online projects with no need to make choices that would leave materials out. This, on the one hand, gives Finnish medievalists comprehensive tools to work with and, on the other, potentially allows for relatively quick progress in (further) developing the services.

In our presentation, we will compare the different Finnish projects, highlight opportunities for international co-operation, and discuss choices (e.g. selecting metadata models) that could best support collaboration between different services and projects.

Critical Play, Hybrid Design and the Performance of Cultural Heritage Game/Stories

Lissa Holloway-Attaway, Associate Professor, School of Informatics, University of Skövde (Skövde, Sweden)

In my talk, I propose to discuss the critical relationship between games designed and developed for cultural heritage and emergent Digital Humanities (DH) initiatives that focus on (re-)inscribing and reflecting on the shifting boundaries of human agency and its attendant relations. In particular, I will highlight theoretical and practical humanistic models that are conceived in tension with more computational emphases and influences. I examine how digital heritage games move us from an understanding of digital humanities as a "tool" or "text" oriented discipline to one where we identify critical practices that actively engage and promote convergent, hybrid and ontologically complex techno-human subjects to enrich our field of inquiry as DH scholars.

Drawing on a few concrete examples from heritage games developed within my university research group and in the heritage design network I co-founded (the Designing Digital Heritage Network), I will outline the important connections between DH and games for heritage. I will focus in particular on a local heritage project created as a transmedial children's books series that re-imagines Nordic folktales for local heritage sites and includes a digital Augmented reality tool, as well as another AR experience designed to complement a book and its attendant archival source materials about the development of the Brooklyn Bridge in the 19th C. These examples will enable me to give an overview of core principles about digital heritage games and align them with emerging hybrid DH initiatives to exemplify future development and research directions. This includes research around new digital literacies, collaborative and co-design approaches (with users) and experimental storytelling and narrative approaches for locative engagement in open-world settings, dependent on input from user/visitors.

Exploring principles such as embodiment, affect, and performativity, and analyzing transmedial storytelling and mixed reality games designed for heritage settings, I argue these games are an exemplary medium for enriching interdisciplinary digital humanities practices using methods currently called upon by recent DH scholarship. In these fully hybrid contexts where human/technology boundaries are richly intermingled, we recognize the importance of theoretical approaches for interpretation that are performative, not mechanistic (Drucker, in Gold, 2011): That is we look at emergent experiences, driven by human intervention, not affirmed by technological development and technical interface affordances. Such hybridity, driven by human/humanities approaches is explored more fully, for example, in Digital_Humanities by Burdick et al (2012) and by N. Katherine Hayles in How We Think: Digital Media and Contemporary Technogenesis (2012). Currently, Hayles and others, like Matthew Gold (2012) offer frameworks for more interdisciplinary Digital Humanities methods (including Comparative Media and Culture Studies approaches) that are richly informed by investigations into the changing role and function of the user of technologies and media and the human/social contexts for use. Hayles, for example, explicitly claims that in Digital Humanities humans " think, through, with, and alongside media" (1). In essence, our thinking and being, our digitization and our human-ness are mutually productive and intertwined. Furthermore, we are multisensory in our access to knowing and we develop an understanding of the physical world in new ways that reorient our agencies and affects, redistributing them for other encounters with cultural and digital/material objects that are now ubiquitous and normalized.

Ross Parry, museum studies scholar, supports a similar model for inquiry and future advancement, based on the premise that digital tool use is now fully implemented and accepted in museum contexts, and so now we must deepen and develop our inquiries and practice (Parry, 2013). He claims that digital technologies have become normative in museums and that currently we find ourselves, then, in the age of the postdigital. Here critical scrutiny is key and necessary to mark this advanced state of change. For Parry this is an opportune, yet delicate juncture that requires a radical deepening of our understanding of the museums' relationship to digital tools:

> Postdigitality in the museum necessitates a rethinking of upon what museological and digital heritage research is predicated and on how its inquiry progresses. Plainly put, we have a space now (a duty even) to reframe our intellectual inquiry of digital in the museum to accommodate the postdigital condition. [Parry, 36]

In line with Parry, and with current DH calls for development, I suggest that we should now focus on the contextualized practices in which these technologies will inevitably engage designers and users and promote robust theoretical and practical applications. To that end, I argue that games, and in particular digital games designed for heritage experiences, are unique training grounds for imagining postdigital future development. They could provide rich contexts for DH scholars working to deepen their understanding of performative and active interventions and intra-actions beyond texts and tools. As digital games have been adopted and ubiquitously assimilated in museums and heritage sites, we have opportunities to study experiences of users as they performatively engage postdigital museum sites through rich forms of hybrid play. In such games, nuanced forms of interdisciplinary communication and storytelling happen in deeply integrated and embedded user/technology relationships. In heritage settings, interpretation is key to understanding histories from multiple user-driven perspectives, and it happens in acts of dynamic emergence, not as the result of mechanistic affordance. As such DH designers and developers have much to learn from a rich body of games and heritage research, particularly that focused on critical and rhetorical design for play, Mixed Reality (MR) approaches and users' bodies as integral to narrative design (Anderson et. al, 2010; Bogost, 2010; Flanagan, 2013; Mortara et. al, 2014; Rouse et. al, 2015; Sicart, 2011).

Additionally, MR provides a uniquely layered approach working across physical and digital artifacts and spaces, encouraging polysemic experiences that can support curators' and historians' desires to tell ever more complex and connected stories for museum and heritage site visitors, even involving visitors' own voices in new ways. In combination, critical game design approaches and MR technologies, within the museum context, help re-center historical experience on the visitor's body, voice, and agency, shifting emphasis away from material objects, also seen as static texts or sites for one-way, broadcast information. Re-centering the design on users' embodied experience with critical play in mind, and in MR settings, offers rich scholarship for DH studies and provides a variety of heritage, museum, entertainment, and participatory design examples to enrich the field of study for open, future and forward thinking.

References

Anderson, E. F., McLoughlin, L., Liarokapis, F., Peters, C., Petridis, P., de Freitas, S.

Developing Serious Games for Cultural Heritage: A State-of-the-Art Review. In: Virtual Reality 14 (4). (2010)

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., Schnapp, J. Digital_Humanities. MIT Press, Cambridge, MA (2012)

Bogost, I. Persuasive Games: The Expressive Power of Videogames. MIT Press, Cambridge MA (2010)

Flanagan, M. Critical Play: Radical Game Design. MIT Press, Cambridge MA (2013)

Gold, M. K. Debates in the Digital Humanities. University of Minnesota Press, Minneapolis, MN (2012)

Hayles, K. N. How We Think: Digital Media and Contemporary Technogenesis. Chicago, University of Chicago Press, Chicago Il (2012)

Parry, R. The End of the Beginning: Normativity in the Postdigital Museum. In: Museum Worlds: Advances in Research, vol. 1, pgs. 24-39. Berghahn Books (2013)

Mortara, M., Catalano, C.E., Bellotti, F., Fiucci, G., Houry-Panchetti, M., Panagiotis, P. Learning Cultural Heritage by Serious Games. In: Journal of Cultural Hertiage, vol. 15, no. 3, pp. 318-325. (2014)

Rouse, R., Engberg, M., JafariNaimi, N., Bolter, J. D. (Guest Eds.) Special Section:

Understanding Mixed Reality. In: Digital Creativity, vol. 26, issue 3-4, pp. 175-227. (2015)

Sicart, M. The Ethics of Computer Games. MIT Press, Cambridge MA (2011)

Digital Humanities in the Nordic Countries, 3rd Conference, 7–9 March 2018, Helsinki

Matti La Mela (History of Industrialization & Innovation group, Aalto University, matti.lamela@aalto.fi)

Paper proposal, short presentation (10 + 5)

Digitised newspapers and the geography of the nineteenth-century ”lingonberry rush” in Finland

This paper uses digitized newspaper data for analysing practices of nature use. In the late nineteenth century, a “lingonberry rush” developed in Sweden and Finland due to the growing foreign demand and exports of lingonberries. The Finnish newspapers followed carefully the events in the neighbouring Sweden, and reported on their pages about the export tons and the economic potential this red gold could have for Finland. The paper is interested in the geography of this “lingonberry rush” and explores how the imprecise geographic information about berry picking can be gathered and structured from the digitized newspapers.

The paper proceeds in two steps. First, it discusses the range of geographic/spatial information included in newspaper data and its potential for historical research. The data used is the historical digitised Finnish newspapers collection of the National Library of Finland (openly accessible until 1929). The spatial information opens different perspectives on the “pieces of news” and derives from both the newspaper metadata and the textual content (where the spatial information is of locational and relational kind). Second, the paper extracts the above-presented geographic information from a semi-automatically generated corpus about lingonberries. Semi-automatic work enables to study closely the qualities and suitability of the data, and, thus, to discuss and exemplify the challenges (OCR quality and its improvement) and possibilities related to named-entity recognition and text reuse detection (the geography of originality and longer chains of news).

This geospatial analysis adds to the reinterpretation of the history of the Nordic allemansrätten, a tradition of public access to nature, which allows everyone to pick wild berries today. In contrast to earlier scholarship that has highlighted the Nordic cultural context, this geospatial analysis draws attention to economic impulses and commercial imagination. The circulation of commercial news on lingonberries (especially about Sweden and Germany) enforced the idea of wild berries as a commodity, and ultimately facilitated to portray the common wild berries as an openly accessible resource.

Identifying poetry based on library catalogue metadata, Hege Roivainen

Changes in printing reflect historical turning points: what has been printed, when, where and by whom are all derivatives of contemporary events and situations. Excessive need for war propaganda brings out more pamphlets from the printing presses; university towns produce dissertations, which scientific development can be deduced from; and strict oppression and censorship might allow only religious publications by government-approved publishers. National library metadata catalogues have been used as sources for studying these turning points.

The traditional way of using national library metadata catalogues in research is simply for finding the location of physical reference books in the library (Altick & Fenstermaker, 1992, 155-182). This kind of research is qualitative in nature - based on close reading and requiring a profound knowledge of the subject matter. While library catalogues may be exploited in this manner for first appearances of various phenomena, the extent of new innovations will not be verified. For example, close readings do not reveal, at least easily, the timeline of Luther's publications, or what portion of books actually were octavo-sized, and when the increase in this format occurred. National library catalogues often contain, more or less, complete records of practically everything published in a certain country or linguistic area in a certain time period. Metadata included often covers information about: physical properties, such as book size and page counts; publisher details; publication places; and so forth. This has made them ideal sources for researchers interested in quantitative analysis and computational approaches aimed at connecting historical turning points and measurable changes in printing. For example, the impact of a new concept can be measured against the amount of re-publications, or the spread of the book, which introduced a new idea (Lahti, Ilomäki & Tolonen, 2015). What is more, linking library metadata to the full text of the books has made it possible to analyse the changes in the usage of words in massive corpora, while still limiting analysis to relevant books (Kanner et al., 2017).

In all these cases, however, computational methods work better the more complete the corpus is, and in the case of library catalogues, there are often deficiencies in annotations. Tthe reasons for this are varied: annotating resources might have been limited, or the annotation rules may have varied between different libraries in cases where catalogues have been amalgamated, or rules could have simply changed. (Karian, 2011).

One area which could be particularly important for researchers is genre. The genre field could be used to restrict the corpus to contain every one of the books that are needed and nothing more. From this subset, then, there would be the possibility of drawing timelines or graphs based on bibliographic metadata, or in the case of full texts, the language or contents of a complete corpus could be analysed. Yet, despite the potential significance of genre information, the field is often unannotated.

My research is a case study which aims at identifying works of poetry in the English Short Title Catalogue (ESTC)[1]. Poetry was chosen, first, because it is a fairly common genre in the ESTC, and second, it is a point of interest for literary researchers. A nearly complete subset of English poetry would allow for large-scale quantitative poetry analysis. With regard to the ESTC: the catalogue contains nearly half a million records of books printed either in English or in Great Britain in the early modern era. Genre information, however, only exists for approximately one fourth of the records.[2]

[1] I had a data dump of ESTC at my disposal for research purposes, so I did not use the online version (http://estc.bl.uk/).
[2] Each record in the ESTC may have multiple genre definitions which are based on the Rare Books and Manuscripts Section (RBMS) genre descriptions. (RBMS, 1991). Also, the depth of categorization varies from micro- to macro-level, which makes assembling relevant subsets difficult.

Rather than relying solely on annotations, a model for machine learning can be taught that uses the more complete aspects of the records, and then genre information deduced from it. In this case, my solution for handling incomplete genre information was turning the genre detection task into a binary classification problem. Each genre value in the catalogue genre labels is judged either as belonging to poetry or non-poetry based on "family resemblance."[3] To this end, I used existing RBMS genre descriptions focusing on versified text, excluding music or performance focused genres such as 'Opera' and 'Plays', but including 'Songs' and 'Ballads'[4]. In the ESTC there are almost 14,000 records which have exactly 'Poems' as a genre, but over 35,000 records containing a genre label describing poetry in this manner. I then took a subset of all the records where the genre field was annotated, and divided it further into training and testing sets. I then trained and tested several different models, which were mostly extracted from the title and subtitle fields. Each of the models was based on a different aspect of the records (such as bag-of-words of the most common words in poetry book titles, part-of-speech tags, or topics). From these models I created a superset of best performing features, which I again tested with the same material. The resulting model performed well: despite the compactness of the metadata, poetry books could be tracked from the test data with a precision over 95%.

I then used the superset of features to seek poetry books without the annotated genre field. This resulted in identifying 13,000 unannotated poetry books from nearly 340,000 records without genre annotation. A sample of one hundred of both poetry and non-poetry findings to manually estimate the correctness of the predictions showed precision of 73%. These results were hampered by an annotation bias in the catalogue. The bias seems to come from two factors: first, one-paged broadside poems have largely had their genre correctly identified in the metadata. Secondly, these works of poetry have been identified as two-page works. Therefore, there is a statistical weighting issue in the training material which identifies two-page works as poetry. In addition, when dealing with works longer than a page, poetry books seem to be annotated more comprehensively than non-poetry. Therefore, by excluding broadsheets from my samples precision rose to 98%.

My research strongly suggest that semi-supervised learning can be applied to library catalogues to fill in missing annotations, but this requires close attention to avoid possible pitfalls.

References

Richard D. Altick and John J. Fenstermaker. *The Art of Literary Research*. Fourth Edition. Norton ; New York. 1993. ISBN 0-393-96240-7.

ESTC. English Short Title Catalogue. URL http://estc.bl.uk/. Accessed 2018-02-05.

Alastair Fowler. *Kinds of Literature. An Introduction to the Theory of Genres and Modes.* Clarendon Press ; Oxford. 1982. ISBN 0-19-812812-6.

John Frow. *Genre*. The New Critical Idiom. Routledge London ; New York, 2006. ISBN 0-415-28063-X.

[3] In genre theory, the concept of family resemblance is often applied to define genres (for example Wolf, 2015; Fowler, 1982, 40-44). The basic idea behind family resemblance is that not all the genre markers have to be present for a text to be considered as belonging to the genre. This is especially emphasized in library catalogues, where the full texts are not available.

[4] There exists a conventionalized main genre division between prose, drama and lyric from the 17th century (Frow, 2006, 59). I have adapted this division.

Antti Kanner, Jani Marjanen, Ville Vaara, Hege Roivainen, Viivi Lähteenoja, Laura Tarkka-Robinson, Eetu Mäkelä, Leo Lahti, and Mikko Tolonen. OCTAVO – Analysing Early Modern Public Communication [poster]. Presented in Digital Humanities at Oxford Summer School. 2017. URL https://comhis.github.io/posters/octavo/. Accessed 2018-02-05.

Stephen Karian. The limitations and possibilities of the estc. *The age of Johnson*, 21:283–297, 2011.

Leo Lahti, Niko Ilomäki, and Mikko Tolonen. A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800. *LIBER Quarterly*, 25(2):87, Dec 2015. ISSN 2213-056X. doi: https://doi.org/10.18352/lq.10112.

RBMS. Genre Terms : A Thesaurus for Use in Rare Book and Special Collections Cataloguing, 1991. URL https://rbms.info/vocabularies/genre/alphabetical_list.htm. Accessed 2018-02-05.

Werner Wolf. The Lyric: Problems of Definition and a Proposal for Reconceptualisation. In *Theory into Poetry. New Approaches to the Lyric*. Edited by Eva Müller-Zettelmann and Margarete Rubik. Pages 21-56. Rodopi ; Amsterdam. 2005. ISBN 90-420-1906-9.

Semantic Annotation of Cultural Heritage Content

Uldis Bojārs[1,2] and Anita Rašmane[1]

[1] National Library of Latvia, Mūkusalas iela 3, Rīga, LV-1423, Latvia
[2] Faculty of Computing, University of Latvia,
Raiņa bulvāris 19, Rīga, LV-1459, Latvia
[uldis.bojars,anita.rasmane]@lnb.lv

Abstract.

This talk focuses on semantic annotation of textual content and on annotation requirements that emerge from the needs of cultural heritage annotation projects. The information presented here is based on two text annotation case studies at the National Library of Latvia and was generalised to be applicable to a wider range of annotation projects.

The case studies examined in this work are (1) correspondence (letters) from the late 19th century between two of the most famous Latvian poets Aspazija and Rainis, and (2) a corpus of parliamentary transcripts that document the first four parliament terms in Latvian history (1922-1934). A pilot project for the personal correspondence annotation case study is available online[1]. It precedes the annotation model described here and was focused on representing annotated and interlinked personal correspondence as Linked Data [1].

The first half of the talk focuses on the annotation requirements. While there were existing annotation tools and data formats available we decided to start with a clean slate and collect the requirements based on the actual annotation needs of the case studies. We propose a semantic annotation model based on these requirements. The annotation model and related requirements are described in more detail in [2].

The model includes support for three core types of annotations - *simple annotations* that may link to named entities, *structural annotations* that mark up portions of the document that have a special meaning within the context of the document (e.g. interruptions and exclamations in a talk at the parliament) and *composite annotations* for more complex use cases (e.g. for representing voting results).

A distinguishing feature of our approach is the introduction of the Entity database that maintains information about the entities referenced from annotations. This allows experts to build a knowledge base about the entities referenced from annotations while annotating documents. This entity information could evolve as

[1] Main page: http://runa.lnb.lv/ – Annotation example: http://runa.lnb.lv/61582/

the annotation task progresses. For example, they may create a record for an entity that needs further research (with comments about what is known about the entity and what is not) which can be extended with additional information (such as identifiers for the entity in other authoritative data sources) when it becomes available.

In the second half of the talk we will present a web-based semantic annotation tool prototype that was developed based on this annotation model and requirements. It allows users to import text documents, create annotations and reference the named entities mentioned in these documents. Information about the entities referenced from annotations is maintained in a dedicated Entity database that supports links between entities and can point to additional information about these entities (e.g. to Linked Open Data resources such as VIAF). Information about these entities is published as Linked Data. Annotated documents may be exported (along with annotation and entity information) in a number of representations including a standalone web view.

References

1. Bojars, U. Case Study: Towards a Linked Digital Collection of Latvian Cultural Heritage. Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe 2016), pp.21-26.
2. Bojars, U., Rasmane, A., Zogla, A. The Requirements for Semantic Annotation of Cultural Heritage Content. Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017. CEUR Workshop Proceedings, vol. 2014. URL: http://ceur-ws.org/Vol-2014/

Layers of History in Digital Games (revised abstract)

Derek Fewster

University of Helsinki ; derek.fewster@helsinki.fi

The past five years have seen a huge increase in historical games studies. Quite a few texts have tried to approach how history is presented and used in games, considering everything from philosophical points to more practical views related to historical culture and the many manifestations of heritage politics.[i] The popularity of recent games like Assassin's Creed, The Witcher and Elder Scrolls also manifests the current importance of deconstructing the messages and choices the games and the game designers present. The impact of especially digital games on the modern understanding of history, and the general idea of time, narrative and change, is yet to be seen in its full effect.

The short paper at hand is an attempt to structure the many layers or horizons of historicity in digital games, into a single taxonomic system for researchers. The suggestion considers the various consciousnesses of time and narrative models modern games work with. Several distinct horizons of time, both of design and of the related real life, are interwoven to form the end product. The field of historical game studies could find this tool quite useful, in its urgent need to systematize how digital culture is reshaping our minds and pasts.

The model considers aspects like memory culture, uses of period art and apocalyptic events, narrative structures, in-game events and real world discourses as parts of how a perception of time and history is created or adapted, interestingly enough often along the same structures if there is any narrative developments within the game. The suggested "layering of time" is applicable on a wide scale of digital games, ranging from Fallout and Dishonored into educational simulations.

In brief, a structuring along six concepts of adaptable but necessary "horizons" is suggested, here with Mass Effect as an example:

Layers	Uses of History or Time	Mass Effect 1-3
Obligating horizon	-> Memory culture, History [in game]	Cycles of evolution
Authenticity horizon	-> Design choices, styles, period art	Rather near future
Apocalyptic horizon	-> Mother of events, reason or leading to the following events	**Arrival of the Reapers (2183)**
Protagonist or Plot horizon	Personal narrative start, possible trauma	Birth 2154, Commoner, Lone survivor
Present horizon	**-> Game events**	2183-, Arrival of Reapers 2186
Related horizon	-> Real world discourses, ideas, choices, morality - right now	**Racism, Essence of humanity, Gender**

i See for example Jason Begy, ”Board Games and the Construction of Cultural Memory” in *Games and Culture* Vol 12, Issue 7-8 (Sage Journals, 2017); Adam Chapman, *Digital Games as History: How Videogmes Represent the Past and Offer Access to Historical Practice* (Routledge, 2016); Dawn Spring, “Gaming history: computer and video games as historical scholarship” in *Rethinking History: The Journal of Theory and Practice*, Vol 19 (2015); the anthology “Gamevironments of the Past”, ed. Derek Fewster & Ylva Grufstedt, *Gamevironments* Issue 5 (2016) & two older articles by Jeremy Antley in *Journal of Digital Humanities*, Vol. 1, No. 2 (2012) on “Going Beyond the Textual in History” & ”Games and Historical Narratives” and the references in all of these works.

Digital Humanities Meet Literary Studies: Challenges Facing Estonian Scholarship

Piret Viires (Tallinn University), Marin Laak (Estonian Literary Museum)

(Contribution ID : 266)

The aim of this paper is to check out the limits and possibilities of DH as a concept and to determine their suitability for literary research in the digital age. Our discussion is rooted, first, in our long-time experience in working with digital models of the Estonian literary history narrative, which are based on literary sources and cultural heritage and, second, in the synchronous study of digitally born new literary forms.

Having dealt with these subjects already for twenty years, we would now attempt to fit our previous research into the context of DH and start to examine the relations between DH and literary studies. Our most general question in this discussion is whether DH can essentially be considered as a tool/method for literary research or is it an entirely new approach in literary studies.

The concept of DH has found active use only in the recent decade, covering a wide range of different areas. In literary research, the adoption of the possibilities offered by DH has been slower than in other fields, mostly because the application of traditional tools of literary research does not require the study and processing of massive amounts of texts which should be done with the help of computer technology. Although the field of DH is broad, in Estonia this concept has rather been used as a synonym for methods of quantitative computer analysis: linguists have already for decades used it very productively in analysing text corpora and computer linguistics has developed into an independent discipline.

In recent years, the application of DH as a method of computerised analysis and the extensive digitisation of literary texts, making them accessible as open data and organising them into large text corpora, have made the relations between literature and information technology a hot topic.

Massive digitisation of cultural heritage – archives and books – was begun at Estonian memory institutions in the early 2000s, at the same time posing a fundamental question: how to select what is important among millions of different books?

Several literary historical projects were initiated even earlier as to offer alternatives to digitisation projects. In Estonia, research on the electronic new media and the application of digital technology in the field of literary studies can be traced back to the second half of the 1990s. The analysis of social, cultural and creative effect (see Schreibman, Siemens, Unsworth 2016: xvii-xviii), as well as constant cooperation with social sciences in the research of the Internet usage have played an important role in Estonian literary studies.

Up to the present, our research has followed three main directions: 1) new forms of literary genres in the electronic environment; 2) digitisation of earlier literature and the creation of a digital bookshelf; 3) development of a new model of the literary historical narrative for applying in the digital environment and the creation of interactive information environments.

1) Digitally born literature and the appearance of other new forms of art have been examined in Estonia since 1996, when the first hypertextual poems were created (e.g. "Trepp" by Hasso Krull), followed by more complex works of digital literature combining different media (text, video, sound, image). Our question is how could the digitally born literature (both "electronic literature" and the literature born on social media, e.g. twitterature, Alt Lit) be fitted into the context of DH? Which are the specific features of digital literature, which are its accompanying effects and how has the role of the reader as the recipient changed in the digital environment? What is the meaning of the concept of hybridity regarding this type of literature? How far can the margin of digital literature be extended to be still called literature? However, the central question is whether DH is able to offer a new framework for analysing digital literature, while so far, highly varied approaches have been used, e.g. the studies of hypertextuality, cybertexts, trans- and intermediality, participation culture, etc.

2) The project for digitising earlier Estonian literature and creating a digital bookshelf "EEVA. The Text Corpus of Earlier Estonian Literature" was created at the University of Tartu in 2002; it makes accessible mostly the works of Baltic German writers. The project is based on the concept of geocultural literature. Books published in all languages in the old Provinces of Estonia and Livonia in the 16th and 17th centuries were digitised. Biographical information about the authors adds much value to the project, which was based on voluminous traditional literary historical research and resulted in a digital bookshelf, where the books can be browsed in the image format https://utlib.ut.ee/eeva/

3) Three large-scale projects for digital representation of Estonian literary history were initiated during the years 1997–2007, with the objective of developing a model of the new literary historical narrative for applying in the digital environment and creating new interactive information environments. These activities were based on the visualisation of the networks of relations between literary works.
   a) The Estonian Literary Museum carried out an Estonian Tiger Leap project "ERNI. Estonian Literary History in texts 1924-1925" in 1997-2001. The project tested the method of reception aesthetics in representing the Estonian literary history of the 1920s. Its objective was to use a relatively limited amount of well-studied material in testing a new type of literary historical narrative and it was based on the visualisation of the network of relations between literary texts and metatexts in the form of hypertext. http://www2.kirmus.ee/erni/erni.html

   b) At the University of Tartu, the project "The Estonian National Epic *The Kalevipoeg*" was developed within the framework of the project CULTOS (Cultural Tools of Learning: Tools and Services *IST-2000-28134*) in 2001–2003. Again, it was a project for visualising literary relationships, requiring the knowledge of the source text and intertexts and reproducing them in the form of a network of intertextual relations. A new software was developed which enabled to explain the relations between the

epic and new multimedia artefacts (literature, art, music, theatre, TV broadcasts, comics, caricatures, etc.).

c) The project "Kreutzwald's Century: the Estonian Cultural History Web" (Kreutzwaldi sajand: eesti kultuurilooline veeb) (in progress) was created at the Estonian Literary Museum in 2004 with the objective of modelling and representing a new narrative of literary history. This was a hybrid project which synthesised the study of the classical narrative of literary history, the needs of the user of the digital new media theory, and the development of digital resources for memory institutions. The underlying idea of the project was to make all the works of fiction of one author, as well as their biography, archival sources, etc., dynamically visible for the reader on an interactive time axis. The scope of the project required the inclusion of all works of fiction published in 1850-1918. Special software was developed for creating the networks of relations between the authors and their works. Kreutzwald's century visualises literary history through the texts and cultural heritage. By now, this project has been in progress already for 15 years, and a voluminous corpus of literary texts has been created. Based on the works of literature digitised for this long-term project, a three-year project "Estonian Literary Classics", carried out with the support of the Estonian Ministry of Culture, made the digitised works accessible for readers in the form of e-books (e-pub format) free of charge. http://kreutzwald.kirmus.ee/

These projects and the research results will now need a further conceptualisation in the context of DH. The projects were not simple digitisation projects, but each of them had their own unique literary theoretical concept, focussing on the creation of relations and the visualisation of textual networks without using quantitative methods. Visualisation of networks was preceded by long-term research using traditional methods, which were based on different theoretical frameworks, such as reception aesthetics, semiotics and narratology. In addition to these traditional methods, the studies of new media and communication (ICT) were also taken into account. At the same time, Estonian projects were not based on single texts, but on literature as a system: the focus was on visualising the sources of literary history and revealing their interrelations. The most important next question is to find ways for further development of the theoretical conceptions of such research and to ask if it is possible to apply the practices and tools of DH for creating new aspects for literary theory.

Thus, one of the aims of our paper is the mapping of different directions, practices and applications of DH in the literary theory of today. One topical question for instance is how to bridge the gap between the research possibilities offered by the present day DH, and the ever increasing resources of texts, produced by memory institutions. Here we encounter several problems. Literary scholars are used to working with texts, analysing them as undivided works of poetry, prose or drama. Using of DH methods requires the treating of literary works or texts as data, which can be analysed and processed with computer programmes (data mining, using visualisation tools, etc.). These activities require the posing of new and totally different research questions in literary studies.

Susan Schreibman, Ray Siemens and John Unsworth, the editors of the book *A New Companion to Digital Humanities* (2016), discuss the problems of DH and point out in their Foreword that it is still questioned whether DH should be considered a separate discipline or, rather, a set of different interlinked methods. In our paper we emphasise the diversity of DH as an academic field of research and talk about other possibilities it can offer for literary research in addition to computational analyses of texts.

However, the final and so far open-ended question is whether DH is essentially a tool for literary research, or is it an entirely new research approach? We need to know whether DH will essentially change literary research because its research objects are not traditional works of literature but digital works which are located in a digital environment. We ask whether DH will form a basis and offer opportunities for a new literary theory? Or perhaps, the artistic essence of a work of art remains the same irrespective of its environment and DH will simply offer a technical package for studying this artistic essence? A further discussion is needed for finding answers to these challenging questions.

Contact:

Piret Viires piret.viires@tlu.ee

Marin Laak marin.laak@kirmus.ee

The plague transformed: *City of Hunger* as mutation of narrative and form

Jennifer J Dellner, Ph.D.

Ocean County College, Toms River, NJ 08754, USA

This short paper proposes and argues the hypothesis that Minna Sundberg's interactive game in development, *City of Hunger*, an offshoot or spin-off of her well respected digital comic, *Stand Still Stay Silent,* can be understood in terms of the ecology of the comic as a mutation of it; as such, her appropriation of a classic game genre and her storyline's emphasis on the mechanical over the natural suggest promising avenues for understanding the uses of interactivity in the interpretation of narrative.

At once a diasporic and plague narrative, Minna Sundberg's web comic *Stand Still. Stay Silent* presents a particularly Nordic post-apocalyptic tale of exploration and discovery, one that elaborates a future world of Scandinavian folkloric creatures and monsters encountered by a group of young explorers who have left the safety of plague-free safe zones in order to discover what is to be found in the rest of the world. The supernatural elements (monsters, grotesquely mutated animals, and souls unable to escape the world) associated with the world of Nordic folklore are also aligned with the plague. A controlling idea throughout is that "the illness" is ecological deformation: writing of "Beasts, Trolls, and Giants," the narrator explains, "They are a shadow of our past, a distorted echo of what once there was." Avoiding the shadow of the past and the monstrosities it has produced is a powerful theme, carrying an implied social critique that deserves examination. In an environment divided into safe areas and the Silent World, the first rule beyond safe zones is avoidance: "do not run or call for help but stand still and stay silent. It might go away" (Sundberg, 2013: 68).

In similar fashion, however, to what Dadey argues is a strategy found in other comics about illness, *Stand Still Stay Silent* manifests the plague-riven world in visual terms that challenge the verbal narratives the characters have heard about it, presenting seeing and vision as experiential knowledge that upends and exceeds received wisdom. While Christensen insists that *Stand Still Stay Silent*'s themes of isolation and fear of contagion fit formulaic plague narrative perfectly, e.g. when the people flee toward isolated places, "huts in the mountains or …. Iceland" (22), completely free of the plague, this is only in the exposition of the story world. The governing principle of the comic, both structural and thematic, is transgression.

While the characters often go where they are not supposed to go, a central feature of the social interaction of those on the mission is debate about what to do and how to interpret their experiences. As some only speak Finnish and others only Norwegian or Swedish (this is indicated by a drawing of each country's flag), a few are multi-lingual; in order to communicate, they must cooperate as a group. These moments in the plot invite readers to also engage in debate in the comments feature, offer solutions and plot suggestions in between postings of the new panel(s), a feature Sundberg herself wrote about in her undergraduate thesis. A key means of building a readership, she writes, is to create, via a comments section, a means for readers to build personal connections with the world of the narrative and the author; social media is another (2013:17). A key component for the act of reading here is interactivity, and is enabled by the digital environment of the comic. The capacity for, and the necessity of, interaction is emphasized by both the plot and the digital affordances of the comic, and challenges expectations that the characters and readers might bring concerning the safety of isolation and fear of others.

In *Stand Still Stay Silent,* biological mutation is the root of the ecology; it is the trope of both transformation and death alike. *City of Hunger,* Sundberg's interactive game in development, can, in fact, be read as a mutation of this original world. *City of Hunger* extracts and foregrounds a subtext found in the comic's plague narrative as the basis of its structure and theme: "the illness" as a battle against malevolent forces. In the game, the illness may or may not be gone, but conflict (vs. cooperation) becomes the primary

mode of interaction for characters and reader-players alike. In order to produce the narrative, the reader-player will have to do battle as the characters do.[1] Sundberg herself signals that her new genre is indivisible from the different ecology of the game world's narrative. "*City of Hunger* will be a 2d narrative rpg with a turn-based battle system, mechanically inspired by your older final fantasy games, the Tales of-series and similar classical rpg's." There will be a world of "rogue humans, mechanoids and mysterious alien beings to fight" (2017). While it remains to be seen how the game develops, its emphasis on machine-beings and aliens in a classic game environment (a "shadow of the past") suggests strongly that the use of interactivity within each narrative has an interpretive and not merely performative dimension.

References:

Christensen, Joergen Riber. "The Formula of Plague Narratives." *Akademisk kvarter/Academic Quarter.* 12. Issue: October. (2015): 15-29.

Dadey, Bruce. "Breaking Quarantine: Image, Text, and Disease in *Black Hole*, *Epileptic*, and *Our Cancer Year*." *Imagetext 7.2* (2013): n.p. http://www.english.ufl.edu/imagetext/archives/v7_2/dadey/

Goodbrey, Daniel Merlin. "Game Comics: An Emergent Hybrid Form." *Journal of Graphic Novels and Comics.* Vol. 6:1 (2015): 3-14. http://dx.doi.org/10.1080/21504857.2014.943411

Montfort, Nick. *Twisty Little Passages: An Approach to Interactive Fiction.* Cambridge, MA. MIT Press. (2003).

National Cartoonist's Society. Reuben Awards (2015). http://www.reuben.org/2015/05/reuben-awards-winners-2015/

Sundberg, Minna *City of Hunger: a narrative rpg* Devlog. (2017). http://www.hummingfluff.com/

——. *Nettisarjakuvien: taloudelliset mahdollisuudet.* Graafisen suunnittelun kandidaatin opinnäytetyö. (2013): Aalto yliopisto. https://aaltodoc.aalto.fi/bitstream/handle/123456789/10338/optika_id_785_sundberg_minna_2013.pdf

——. *Stand Still. Stay Silent.* (2013-) http://www.sssscomic.com

[1] Montfort notes that adventure games are not themselves narratives, but "produce narratives when a

Poster: *Elias Lönnrot Letters Online*

Kirsi Keravuori, Maria Niku, Finnish Literature Society

The correspondence of Elias Lönnrot (1802–1884, doctor, philologist and creator of the national epic *Kalevala*) comprises of 2 500 letters or drafts written by Lönnrot and 3 500 letters received. *Elias Lönnrot Letters Online* (http://lonnrot.finlit.fi/omeka/), first published in April 2017, is the conlusion of several decades of research, of transcribing and digitizing letters and of writing commentaries. The online edition is designed not only for those interested in the life and work of Lönnrot himself, but more generally to scholars and general public interested in the work and mentality of the Finnish 19th century nationalistic academic community , their language practices both in Swedish and in Finnish, and in the study of epistolary culture. The rich, versatile correspondence offers source material for research in biography, folklores studies and literary studies; for general history as well as medical history and the history of ideas; for the study of ego documents and networks; and for corpus linguistics and history of language.

As of January 2018, the edition contains about 2000 letters and drafts of letters sent by Elias Lönnrot (1802-1884, doctor, philologist and creator of the national epic *Kalevala*). These are mostly private letters. The official letters, such as the medical reports submitted by Lönnrot in his office as a physician, will be added during 2018. The final stage will involve finding a suitable way of publishing for the approximately 3500 letters that Lönnrot received.

The edition is built on the open-source publishing platform *Omeka*. Each letter and draft of letter is published as facsimile images and an XML/TEI5 file, which contains metadata and transcription. The letters are organised into collections according to recipient, with the exception of for example Lönnrot's family letters, which are published in a single collection. An open text search covers the metadata and transcriptions. This is a faceted search powered by Apache's *Solr* which allows limiting the initial search by collection, date, language, type of document and writing location. In addition, *Omeka*'s own search can be used to find letters based on a handful of metadata fields.

The solutions adopted for the Lönnrot edition differ in some respects from the established practices of digital publishing of manuscripts in the humanities. In particular, the TEI encoding of the transcriptions is lighter than in many other scholarly editions. Lönnrot's own markings – underlinings, additions, deletions – and unclear and indecipherable sections in the texts are encoded, but place and personal names are not. This is partially due to the extensive amount of work such detailed encoding would require, partially because the open text search provides quick and easy access to the same information.

The guiding principle of *Elias Lönnrot Letters* is openness of data. All the data contained in the edition is made openly available.

Firstly, the XML/TEI5 files are available for download, and researchers and other users are free to modify them for their own purposes. The users can download the XML/TEI5 files of all the letters, or of a smaller section such as an individual collection. The feature is also integrated in the open text search, and can be used both for all the results produced by a search and a smaller section of the results limited by one or more facets. Thus, an individual researcher can download the XML files of the letters and study them for example with the linguistic tools provided by the Language Bank of Finland. Similarly, the raw data is available for processing and modifying by those researchers who use and develop digital humanities tools and methods to solve research questions.

Secondly, the letter transcriptions are made available for download as plain text. Data in this format is needed for qualitative analysis tools like *Atlas*. In addition, researchers in humanities do not all need XML files but will benefit from the ability to store relevant data in an easily readable format.

Thirdly, users of the edition can export the statistical data contained in the facet listing of each search result for processing and visualization with tools like Excel. Statistical data like this is significant in handling large masses of data, as it can reveal aspects that would remain hidden when examining individual documents. For example, it may be relevant to a researcher in what era and with whom Lönnrot primarily discussed a given theme. The statistical data of the facet search readily reveals such information, while compiling such statistics by manually going through thousands of letters would be an impossibly long process.

The easy availability of data in *Elias Lönnrot Letters Online* will hopefully foster collaboration and enrich research in general. The SKS is already collaborating with Finn-Clarin and the Language Bank, which have received the XML/TEI5 files. As Lönnrot's letters form an exceptionally large collection of manuscripts written by one hand, a section of the letters together with their transcriptions was given to the international READ project, which is working to develop machine recognition of old handwritten texts. A third collaborating partner is the project "STRATAS – Intefacing structured and unstructured data in sociolinguistic research on language change".

Abstract for poster for DHN Digital Humanities in the Nordic Countries 3rd conference 2018

Presenters:
Project Director: Tiina H. Airaksinen (tiina.h.airaksinen@helsinki.fi, Senior Lecture in Asian Studies
Research Assistant: Anna-Leena Korpijärvi (anna-leena.korpijarvi@helsinki.fi, MA, PhD Candidate

Title: KuKa Digi -project

Abstract:

This poster presents a sample of the Cultural Studies BA program's Digital Leap project called **KuKa Digi**. The Digital Leap is a university wide project that aims to support digitalization in both learning and teaching in the new degree programs at the University of Helsinki. For more information on the University of Helsinki's Digital Leap program, please refer to: http://blogs.helsinki.fi/digiloikka/ . The new Bachelor's Program in Cultural Studies, was among the projects selected for the 2018-2019 round of the Digital Leap. The primary goal of the KuKa Digi project is to produce meaningful digital material for both teaching and learning purposes. The KuKa Digi project aims to develop the program's courses, learning environments and materials into a more digital direction. Another goal of the project is to produce an introductory MOOC –course on Cultural Studies for university students, as well as students studying for their A-levels, who may be planning to apply for the Cultural Studies BA program. Finally, we will write a research article to assess the use of digital environments in teaching and learning processes within Cultural Studies BA program. Kuka Digi –project encourages students and teachers to co-operatively plan digital learning environments that are also useful in building up students' academic portfolio and enhance their working life skills.

The core idea of the project is to create a digital platform or database for teachers, researchers and students in the field of Cultural Studies. Academic networking sites do exist, however they are not without issues. Many of them are either not accessible, or very useful for students, who have not developed their academic careers very far yet. In addition to this, some of these sites are only partially free of charge. The digital platform will act as a place where students, teachers and researchers alike can have the opportunity to network, advertise their expertise and specialization as well as, come into contact with the media, cultural agencies, companies and much more. The general vision for this platform is that it will be user friendly, flexible as well as, act as an "academic Linked In". The database will be available in Finnish, Swedish and English. The database will include the current students, teachers and experts, who are associated with the program. Furthermore, the platform will include a feature called the digital

portfolio. This will be especially useful for our students, as it is intended to be a digital tool with which they can develop their own expertise within the field of Cultural Studies. Finally, the portfolio will act as a digital business card for the students. The Project poster presented at the conference illustrates the ideas and concepts for the platform in more detail.

For more information on the project and its other goals, please refer to the project blog at: http://blogs.helsinki.fi/kuka-digi/

Topic modelling and qualitative textual analysis

Karoliina Isoaho, Daria Gritsenko (University of Helsinki)

The pursuit of big data is transforming qualitative textual analysis—a laborious activity that has conventionally been executed manually by researchers. Access to data of unprecedented scale and scope has created a need to both analyse large data sets efficiently and react to their emergence in a near-real-time manner (Mills, 2017). As a result, research practices are also changing. A growing number of scholars have experimented with using machine learning as the main or complementary method for text analysis. Even if the most audacious assumptions 'on the superior forms of intelligence and erudition' of big data analysis are today critically challenged by qualitative and mixed-method researchers (Mills, 2017: 2), it is imperative for scholars using qualitative methods to consider the role of computational techniques in their research (Janasik, Honkela and Bruun, 2009). Social scientists are especially intrigued by the potential of topic modelling (TM), a machine learning method for big data analysis (Blei, 2012), as a tool for analysis of textual data.

This research contributes to a critical discussion in social science methodologies: how topic modeling can concretely be incorporated into existing processes of qualitative textual analysis and interpretation. Some recent studies paid attention to the methodological dimensions of TM vis-à-vis textual analysis. However, these developments remain sporadic, exemplifying a need for a systematic account of the conditions under which TM can be useful for social scientists engaged in textual analysis. This paper builds upon the existing discussions, and takes a step further by comparing the assumptions, analytical procedures and conventional usage of qualitative textual analysis methods and TM. Our findings show that for content and classification methods, embedding TM into research design can partially and, arguably, in some cases fully automate the analysis. Discourse and representation methods can be augmented with TM in sequential mixed-method research design.

We outline avenues for TM both in embedded and sequential mixed-method research design. This is in line with previous work on mixed-method research that has challenged the traditional assumption of there being a clear division between qualitative and quantitative methods. Scholarly capacity to craft a robust research design depends on researchers' familiarity with specific

techniques, their epistemological assumptions, and good knowledge of the phenomena that are being investigated to facilitate the substantial interpretation of the results. We expect this research to help identify and address the critical points, thereby assisting researchers in the development of novel mixed-method designs that unlock the potential of TM in qualitative textual analysis without compromising methodological robustness.

Blei, D. M. (2012) 'Probabilistic topic models', Communications of the ACM, 55(4), p. 77.

Janasik, N.,Honkela, T. and Bruun, H. (2009) 'Text Mining in Qualitative Research', Organizational Research Methods, 12(3), pp. 436–460.

Mills, K. A. (2017) 'What are the threats and potentials of big data for qualitative research?', Qualitative Research, p. 146879411774346.

Heikki Kokko
University of Tampere
Centre of Excellence in the
History of Experiences (HEX)
email: Heikki.j.kokko@uta.fi
28.2.2018

Local Letters to Newspapers – digital history project

Introduction

The Local Letters to Newspapers is a digital history project of the Academy of Finland Centre of Excellence in the History of Experiences HEX (2018–2025), hosted by University of Tampere. The objective is to make a new kind of digital research material available from the 19th and the early 20th century Finnish society. The aim is to introduce a database of the readers' letters submitted to the Finnish press that could be studied both qualitatively and quantitatively. The database will allow analyzing the 19th and 20th century global reality through a case study of the Finnish society. It will enable a wide range of research topics and open a path to various research approaches, especially the study of human experiences.

Local letters to newspapers

Our focus is on the letters to newspapers. They were the letters sent to the press by the readers of the newspapers. They were usually written in the name of the parish, from which the letter was sent, and the name of the parish was used as a title. Therefore, an individual writer represented a whole parish. That is a reason why in this project these letters are called *local letters*.

These kind of letters told usually about every day local things that had happened in some parish. Usually, the topics included information about yields, the state of health of the inhabitants and curious incidents. However, there were also abstract reasoning on topics like for example, what is society, publicity or individual.

There are a hundreds of thousands of these kind of letters written by both Finnish and Swedish in the Finnish press merely in the 19th century. They covered a large part of the print sheets of the whole Finnish-speaking press especially in the mid-1800s. The great number of the local letters were published because the early journalists of the Finnish-speaking press worked only part-time. That is why they were happy to accept a great number of letters to their readers and publish them. In the 1850s and 1860s, the letters to the editor increased to the significant phenomenon. There is a calculation that in the 1847–65 there were at least 2500 writers of these letters. That is a significant number because for example in the 1847 there were only about 1000 and in the 1860 10000 subscriptions. (Tommila 1988) Therefor the readers wrote a large part of the early Finnish press and indeed these local letters were a reason why people began to subscribe and read more and more newspapers.

There is not accurate knowledge of the social status of the writers of the letters, because many of them used pseudonyms. There is a careful estimation, that 40 percent of the identified writers were either peasants or came from the lower social layer. That means that the letters are a good source for “history-from-below” perspective. Indeed, the local letters are the first a source of this kind in

the Finnish-speaking culture, because almost the whole central administration of Finland worked in Swedish until to the end of the 19th century.

The projects main criteria for the "local letter" are:

1. The letters are sent from the local stage to the publicity that is a "translocal" stage. (Excludes for example the notices of administration)

2. The length of the writing makes an experience visible. (Excludes for example the short news flashes)

Database of local letters - Possibilities for qualitative and quantitative analysis

The Local Letters to Newspapers -digital project is constructing a database with the tools provided by the Digital Collections of National Library of Finland. The Scrapbook -tool in the National library's digital collections website allows collecting the cuttings to the database of the project. When the database is ready, this enables the searches that will cover only the local letters and exclude the other newspaper material. This opens the path to do the qualitative "cross-sections" to the whole press in the different time eras. Therefore, this creates a change to analyse some particular historical phenomenon as well as the language itself, for example conceptual history.

There is also a possibility to approach these local letters quantitatively. The background data of the letters could be converted to Excel -mode. This makes possible to group these local letters by the name of the parishes. This enables the study of the societal distribution of the local letters and thus shed light on the societal significance of this phenomenon. It could be for example studied, from where the letters were sent in a particular era.

The old Article index (collected in 1890–1909)

There is an existing index of these local letters that is collected at the turn of the 20th century. This old index originally included 380 000 little hand-written filing cards that are in the file boxes. A part of the hand-written index was typewritten and microfilmed in the 20th century. In the 21st century, they were digitalized and published online by the national library of Finland. https://digi.kansalliskirjasto.fi/sanomalehti/directory

Unfortunately, this originally hand-written index is not complete. There are many blank spots and repetition on it. The criteria of the picking the articles also changed remarkable during a long almost 20-year period of the collect work. That is why this old index is not reliable enough for the quantitative scientific analysis in particular.

Nevertheless, the benefit of this old index is that there are the descriptions of the local letters to newspapers. These originally hand written descriptions outline the content of every indexed letter. The Local Letters to Newspapers -digital project will combine these descriptions of the old Article index with the new Database of Local Letters with the digital technology. This combines the technologies of the 19th and 21st century and provides a new kind of data to the historical research.

See more information about this old Article index (in Finnish): http://blogs.helsinki.fi/scriptaselecta/2017/07/13/digitaalisten-aineistojen-artikkelihakemiston-historiasta/

Open access & Data management

The project has a plan to make the Local Letters to Newspapers -database fully available online for free. The first release will be the period from the beginning of the Finnish press to the year 1870. In the long perspective, the objective is to continue the database to the first decades of the 20th century.

We believe that in any extent, the database will provide a great material for many kinds of historical study, because it allows to use the Finnish society as a case study for global approaches. In that way, the database will serve disciplines like social history, history from below, conceptual history, cultural history or local history.

There are also plans for crowdsourcing the work. The writers who wrote the local letters often used pseudonyms. That is why the identification of the writers is often difficult. Different researchers have different fragments of information about the people behind these pseudonyms. The plan is to join these pieces of information by introducing the public forum to the web, which allows the debate about local letters and especially the writers of the letters. This crowdsourced knowledge could thus be transferred as an entry to the database of local letters.

Current situation of the project

The project is in its starting phase. We are currently recruiting a research assistant whose task will be to collect the local letters to newspapers to the database. In future, there are plans to use computer and OCR reading in this collecting process.

Lessons Learned from Historical Pandemics. Using crowdsourcing 2.0 and Citizen Science to map the Spanish Flus spatial and social network

By Søren K. Poder MA. In history & Astrid Lykke Birkving, MA in intellectual History

Aarhus City Archvies | Redia a/s

In 1918 the World was struck by the most devastating disease in recorded history - today known as the Spanish Flu. In less than one year nearly two third of world's population came down with influenza. Of which between forty and one hundred million people died.

The Spanish Flu in 1918 did not originated in Spain, but most likely on the North American east coast in February 1918. By the middle of Marts, the influenza had spread to most of the overcrowded American army camps from where it soon was carried to the trenches in France and the rest of the World. This part of the story is well known. In contrast the diffusion of the 1918-pandemic, and the seasonal epidemics for that matter, on the regional and local level is still largely obscure. For instance, an explanation on why epidemics evidently tends to follow significantly different paths in different urban areas that otherwise seems to share a common social, commercial and cultural profile, tend to be more theoretical then based on evidence. For one sole reason – the lack of adequate data.

As part of the incessantly scientific interest in historical epidemics, the purpose of this research project is to identify the social, economic and cultural preconditions that most likely determines a given type of locality's ability to spread or halter an epidemic's hieratical diffusion.

Crowdsourcing 2.0

To meet ends data large amounts of data from a variety of different historical sources as to be collected and linked together. To do this we use traditional crowdsourcing techniques, where volunteers participates in transcribing different historical documents. Death certificates, census, patient charts etc. But just as important does the collected transcription form the base for a text recognition ML module that in time will be able recognize specific entities in a document – persons, placers, diagnoses dates ect.

Triadic closure amplifies homophily in social networks

Aili Asikainen, Gerardo Iñiguez, Kimmo Kaski and Mikko Kivelä

keywords: social network analysis, social dynamics, triadic closure, homophily, agent-based modelling

Much of the structure in social networks can be explained by two seemingly separate network evolution mechanisms [1]: triadic closure and homophily. While it is typical to analyse these mechanisms separately, empirical studies suggest that their dynamic interplay can be responsible for the striking homophily patterns seen in real social networks [2]. By defining a network model with tunable amount of homophily and triadic closure, we find that their interplay produces a myriad of effects such as amplification of latent homophily and memory in social networks (hysteresis). We use empirical network datasets to estimate how much observed homophily could actually be an amplification induced by triadic closure, and have the networks reached a stable state in terms of their homophily. Beyond their role in characterizing the origins of homophily, our results may be useful in determining the processes by which structural constraints and personal preferences determine the shape and evolution of society.

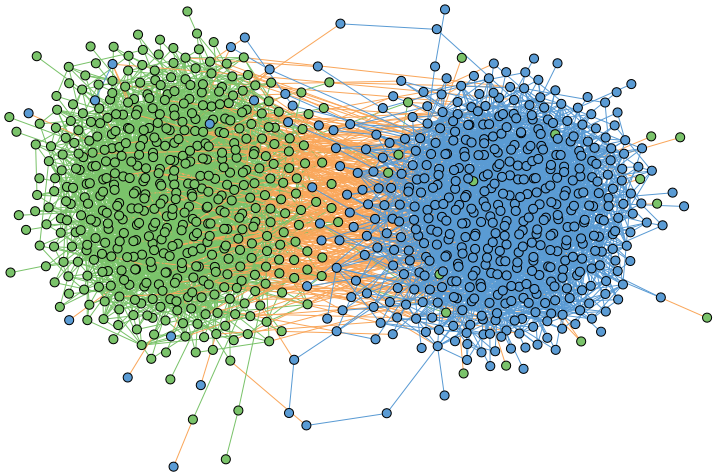


Figure 1: Visualization of a network with large observed homophily produced by the model. The network consists of 964 nodes each of which belongs to one of two groups (green or blue) so that the groups are equal in size. Out of the 5000 links 46.5% are between green nodes, 45.3% are between blue nodes and 8.2% are between a blue and a green node.

References

[1] Toivonen, R. et al., *Soc. Net.* **31**, 240–254 (2009).
[2] Kossinets, G. et al., *Am. J. Sociol.* **115**, 405–450 (2009).

Dissertations from Uppsala University 1602-1855 on the Internet

Dr. Anna Fredriksson, Uppsala University Library

At Uppsala University Library, a long-term project is under way which aims at making the dissertations, that is theses, submitted at Uppsala University in 1602-1855 easy to find and read on the Internet. The work includes metadata production, scanning and OCR processing as well as publication of images of the dissertations in full-text searchable pdf files. So far, approximately 3,000 dissertations have been digitized and made accessible on the Internet via the DiVA portal, Uppsala University's repository for research publications. All in all, there are about 12,000 dissertations of about 20 pages each on average to be scanned. This work is done by hand, due to the age of the material. The project aims to be completed in 2020.

Why did we prioritize dissertations?

Even before the project started, dissertations were valued research material, and the physical dissertations were frequently on loan. Their popularity was primarily due to the fact that generally, studying university dissertations is a great way to study evolvements and changes in society. In the same way as doctoral theses do today, the older dissertations reflect what was going on in the country, at the University, and in the intellectual Western world on the whole at a certain period of time. The great mass of them makes them especially suitable for comparative and longitudinal studies, and provides excellent chances for scholars to find material little used or not used at all in previous research.
Swedish older dissertations including those of today's Finland specifically are also comparatively easy to find. In contrast to many other European libraries with an even longer history, collectors published bibliographies of Swedish dissertations as far back as 250 years ago. Our dissertations are also organized, bound and physically easily accessible. Last year the cataloguing of the Uppsala dissertations was completed according to modern standards in LIBRIS. That made them searchable according to subject and word in title, which was not possible before. All this made the digitization process smoother than that of many other kinds of cultural heritage material. The digital publication of the dissertations naturally made access to them even easier for University staff and students as well as lifelong learners in Sweden and abroad.

How are the dissertations used today?

In actual research today, we see that the material is frequently consulted in all fields of history. Dissertations provide scholars in the fields of history of ideas and history of science with insight into the status of a certain subject matter in Sweden in various periods of time, often in relation to the contemporary discussion on the European continent. The same goes for studies in history of literature and history of religion. Many of the dissertations examine subjects that remain part of the public debate today, and are therefore of interest for scholars in the political and social sciences. The languages of the dissertations are studied by scholars

of Semitic, Classical and Scandinavian languages, and the dissertations often contain the very first editions and translations of certain ancient manuscripts in Arabic and Runic script. There is also a social dimension of the dissertations worthy of attention, as dedications and gratulatory poems in the dissertations mirror social networks in the educated stratum of Sweden in various periods of time. Illustrations in the dissertations were often made by local artists or the students themselves, and the great mass of gratulatory poems mirrors the less well-known side of poetry in early modern Sweden.

Our users

The users of the physical items are primarily university scholars, primarily our own University, but there is also quite a great deal of interest from abroad. Not least from our neighboring country Finland and from the Baltic States, which were for some time within the Swedish realm. Many projects are going on right now which include our dissertations as research material or which have them as their primary source material; Swedish projects as well as international. As Sweden as a part of learned Europe more or less shared the values, objects and methods of the Western academic world as a whole, to study Swedish science and scholarship is to study an important part of Western science and scholarship.
As for who uses our digital dissertations, we in fact do not know. The great majority of the dissertations are written in Latin, as in all countries of Europe and North America, Latin was the vehicle for academic discussion in the early modern age. In the first half of the 19th century, Swedish became more common in the Uppsala dissertations. Among the ones digitized and published so far, a great deal are in Swedish. As for the Latin ones, they too are clearly much used. Although knowledge of Latin is quite unusual in Sweden, foreign scholars in the various fields of history often had Latin as part of their curriculum. Obviously, our users know at least enough Latin to recognize if a passage treats the topic of their interest. They can also identify which documents are important to them and extract the most important information from it. If the document is central, it is possible to hire a translator.

But we believe that we also reach out to the lifelong learners, or the so-called "ordinary people". The older dissertations examine every conceivable subject and they offer pleasant reading even for non-specialists, or people who use the Internet for genealogical research. The full text publication makes the dissertation show up, perhaps unexpectedly, when a person is looking for a certain topic or a certain word. Whoever the users the digital publication of the dissertations has been well received, and far beyond expectations. The first three test years of approximately 2,500 digitized dissertations published resulted in close to one million visits and over 170,000 downloads, i.e. over 4,700 per month. Even if we don't – or perhaps *because* we don't – either offer or demand advanced technologies for the use of these dissertations.

The digital publication and the new possibilities for research

The database in which the dissertations are stored and presented is the same database in which researchers, scholars and students of Uppsala University, and other Swedish universities, too,

currently register their publications with the option to publish them digitally. This clears a path for new possibilities for researchers to become aware of and study the texts. Most importantly, it enables users to find documents in their field, spanning a period of 400 years in one search session. A great deal of the medical terms of diseases and body parts, chemical designations, and, of course, juridical and botanical terms are Latin and the same as were used 400 years ago, and can thus be used for localizing text passages on these topics. But the form of the text can be studied, too. Linguists would find it useful to make quantitative studies of the use of certain words or expressions, or just to find the words of interest for further studies. The usefulness of full-text databases are all known to us. But often one as a user gets either a well-working search system *or* a great mass of important texts, and seldom both. This problem is solved here by the interconnection between the publication database DiVA and the Swedish National Research Library System LIBRIS. The combination makes it possible to use an advanced search system with high functionality, thus reducing the Internet problem of too many irrelevant hits. It gives direct access to the digital full text in DiVA, and the option to order the physical book if the scholar needs to see the original at our library. Not least important, there is qualified staff appointed to care for the system's long-term maintenance and updates, as part of their everyday tasks at the University Library. Also, the library is open for discussion with users.

The practical work within the project and related issues

As part of the digitization project, the images of the text pages are OCR-processed in order to create searchable full-text pdf files. The OCR process gives various results depending on the age and the language of the text. The OCR processing of dissertations in Swedish and Latin from ca. 1800 onwards results in OCR texts with a high degree of accuracy, that is, between 80 and 90 per cent, whereas older dissertations in Latin and in languages written in other alphabets will contain more inaccuracies. On this point we are not satisfied. Almost perfect results when it comes to the OCR-read text, or proof-reading, is a basic requirement for the full use and potential of this material. However, in this respect, we are dependent upon the technology which is available on the market, as this provides the best and safest product. These products were not developed for handling printing types of various sorts and sizes from the 17th and 18th centuries, and the development of these techniques, except when it comes to "Fraktur", is slow or non-existing.

If you want to pursue further studies of the documents, you can download the documents for free to your own computer. There are free programs on the Internet that help you merge several documents of your choice into one document, in order for you to be able to search through a certain mass of text. If you are searching for something very particular, you could of course also perform a word search in Google. One of our wishes for the future is to make it possible for our users to search in several documents of their specific choice at one time, without them having to download the documents to their computer.

So, most important for us today within the dissertation project:

1) Better OCR for older texts
2) Easier ways to search in a large text mass of your own choice.

Future use and collaboration with scholars and researchers

The development of digital techniques for the further use of these texts is a future desideratum. We therefore aim to increase our collaboration with researchers who want to explore new methods to make more out of the texts. However, we always have to take into account the special demands from society when it comes to the work we, as an institute of the state, are conducting – in contrast to the work conducted by e.g. Google Books or research projects with temporary funding.

We are expected to produce both images and metadata of a reasonably high quality – a product that the University can 'stand for'. What we produce should have a lasting value – and ideally be possible to use for centuries to come.

What we produce should be compatible with other existing retrieval systems and library systems. Important, in my opinion, is reliability and citability. A great problem with research on digitally borne material is, in my opinion, that it constantly changes, with respect to both their contents and where to find them. This puts the fundamental principle of modern science, the possibility to control results, out of the running. This is a challenge for Digital Humanities which, with the current pace of development, surely will be solved in the near future.

Normalizing Early English Letters for Neologism Retrieval

Mika Hämäläinen, Tanja Säily and Eetu Mäkelä

Department of Digital Humanities
University of Helsinki

1 Introduction

Our project studies social aspects of innovative vocabulary use in early English letters. In this abstract we describe the current state of our method for detecting neologisms. The problem we are facing at the moment is the fact that our corpus consists of non-normalized text. Therefore, spelling normalization is the first step we need to solve before we can apply automatic methods to the whole corpus.

2 Corpus

We use CEEC (Corpora of Early English Correspondence) [9] as the corpus for our research. The corpus consists of letters ranging from the 15th century to the 19th century and it represents a wide social spectrum, richly documented in the metadata associated with the corpus, including information on e.g. socioeconomic status, gender, age, domicile and the relationship between the writer and recipient.

3 Finding Neologisms

In order to find neologisms, we use the information of the earliest attestation of words recorded in the Oxford English Dictionary (OED) [10]. Each lemma in the OED has information about its attestations, but also variant spelling forms and inflections.

How we proceed in automatically finding neologism candidates is as follows. We get a list of all the individual words in the corpus, and we retrieve their earliest attestation from the OED. If we find a letter where the word has been used before the earliest attestation recorded in the OED, we are dealing with a possible neologism, such as the word *monotonous* in (1), which antedates the first attestation date given in the OED by two years (1774 vs. 1776).

> (1) How I shall accent & express, after having been so long cramped with the **monotonous** impotence of a harpsichord! (Thomas Twining to Charles Burney, 1774; TWINING_017)

The problem, however, is that our corpus consists of texts written in different time periods, which means that there is a wide range of alternative spellings for words. Therefore, a great part of the corpus cannot be directly mapped to the OED.

4 Normalizing with the Existing Methods

Part of the CEEC (from the 16th century onwards) has been normalized with VARD2 [3] in a semi-automated manner; however, the automatic normalization is only applied to sufficiently frequent words, whereas neologisms are often rare words. We take these normalizations and extrapolate them over the whole corpus. We also used MorphAdorner [5] to produce normalizations for the words in the corpus. After this, we compared the newly normalized forms with those in the OED taking into account the variant forms listed in the OED. NLTK's [4] lemmatizer was used to produce lemmas from the normalized inflected forms to map them to the OED. In doing so, we were able to map 65,848 word forms of the corpus to the OED. However, around 85,362 word forms still remain without mapping to the OED.

5 Different Approaches

For the remaining non-normalized words, we have tried a number of different approaches.

- Rules
- SMT
- NMT
- Edit distance, semantics and pronunciation

The simplest one of them is running the hand-written VARD2 normalization rules for the whole corpus. These are simple replacement rules that replace a sequence of characters with another one either in the beginning, end or middle of a word. An example of such a rule is replacing *yes* with *ies* at the end of the word.

We have also trained a statistical machine translation model (with Moses [7]) and a neural machine translation model (with OpenNMT [6]). SMT has previously been used in the normalization task, for example in [11]. Both of the models are character based treating the known non-normalized to normalized word pairs as two languages for the translation model. The language model used for the SMT model is the British National Corpus (BNC) [1].

One more approach we have tried is to compare the non-normalized words to the ones in the BNC by Levenshtein edit distance [8]. This results in long lists of normalization candidates, that we filter further by their semantic similarity, which means comparing the list of two word appearing immediately after and before the non-normalized word and the normalization candidates picking out

the candidates with largest number of shared contextual words. And finally, filtering this list with Soundex pronunciation by edit distance. A similar method [2] has been used in the past for normalization which relied on the semantics and edit distance.

6 The Open Question

The above described methods produce results of varying degrees of success. However, none of them is reliable enough to be trusted above the rest. We are now in a situation in which at least one of the approaches finds the correct normalization most of the time. The next unsolved question is how to pick the correct normalization from the list of alternatives in an accurate way.

Once the normalization has been solved, we are facing another problem which is mapping words to the OED correctly. For example, currently the verb *to moon* is mapped to the noun *mooning* recorded in the OED because it appeared in the present participle form in the corpus. This means that in the future, we have to come up with ways to tackle not only the problem of homonyms, but also the problem of polysemy. A word might have acquired a new meaning in one of our letters, but we cannot detect this word as a neologism candidate, because the word has existed in the language in a different meaning before.

References

1. The British National Corpus, version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium (2007), http://www.natcorp.ox.ac.uk/
2. Amoia, M., Martinez, J.M.: Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In: Proceedings of the 7th workshop on language technology for cultural heritage, social sciences, and humanities. pp. 84–89 (2013)
3. Baron, A., Rayson, P.: VARD2: a tool for dealing with spelling variation in historical corpora (2008)
4. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
5. Burns, P.R.: Morphadorner v2: A java library for the morphological adornment of English language texts. Northwestern University, Evanston, IL (2013)
6. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. ArXiv e-prints
7. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. pp. 177–180. Association for Computational Linguistics (2007)
8. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710 (1966)

9. Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A., Palander-Collin, M.: CEEC, Corpus of Early English Correspondence. Department of Modern Languages, University of Helsinki, http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
10. OED: OED Online. Oxford University Press, http://www.oed.com/
11. Pettersson, E., Megyesi, B., Tiedemann, J.: An SMT approach to automatic annotation of historical text. In: Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18. pp. 54–69. No. 087, Linköping University Electronic Press (2013)

Analysing Swedish Parliamentary Voting Data

Jacobo Rouces, Nina Tahmasebi, Lars Borin, Stian Rødven Eide

We used publicly available data from voting sessions in the Swedish Parliament to represent each member of parliament (MP) as a vector in a space defined by their voting record between the years 2014 and 2017. We then applied matrix factorization techniques that enabled us to find insightful projections of this data. Namely, it allowed the assessment of the level of clustering of MPs according to their party line while at the same time identifying MPs whose voting record is closer to other parties'. It also provided a data-driven multi-dimensional political compass that allows to ascertain similitudes and differences between MPs and political parties. Currently, the axes of the compass are unlabeled and therefore they lack a clear interpretation, but we plan to apply language technology on the parliamentary discussions associated to the voting sessions on order to identify the topics associated to these axis.

Automated Cognate Discovery in the Context of Low-Resource Sami Languages

Eliel Soisalon-Soininen[1] and Mika Hämäläinen[2]

[1] Department of Computer Science
[2] Department of Digital Humanities
University of Helsinki

1 Introduction

The goal of our project is to automatically find candidates for etymologically related words, known as *cognates*, for different Sami languages. At first, we will focus on North Sami, South Sami and Skolt Sami nouns by comparing their inflectional forms with each other. The reason why we look at the inflections is that, in Uralic languages, it is common that there are changes in the word stem when the word is inflected in different cases. When finding cognates, the non-nominative stems might reveal more about a cognate relationship in some cases. For example, the South Sami word for arm, *gïete*, is closer to the partitive of the Finnish word *kättä* than to the nominative form *käsi* of the same word.

The fact that a great deal of previous work already exists related to etymologies of words in different Sami languages [2, 4, 8] provides us with an interesting test bed for developing our automatic methods. The results can easily be validated against databases such as Álgu [1] which incorporates results of different studies in Sami etymology in a machine-readable database.

With the help of a gold corpus, such as Álgu, we can perfect our method to function well in the case of the three aforementioned Sami languages. Later, we can expand the set of languages used to other Uralic languages such as Erzya and Moksha. This is achievable as we are basing our method on the data and tools developed in the Giellatekno infrastructure [11] for Uralic languages. Giellatekno has a harmonized set of tools and dictionaries for around 20 different Uralic languages allowing us to bootstrap more languages into our method.

2 Related Work

In historical linguistics, cognate sets have been traditionally identified using the comparative method, the manual identification of systematic sound correspondences across words in pairs of languages. Along with the rapid increase in digitally available language data, computational approaches to automate this process have become increasingly attractive.

Computationally, automatic cognate identification can be considered a problem of clustering similar strings together, according to pairwise similarity scores given by some distance metric. Another approach to the problem is pairwise

classification of word pairs as cognates or non-cognates. Examples of common distance metrics for string comparison include edit distance, longest common subsequence, and Dice coefficient.

The string edit distance is often used as a baseline for word comparison, measuring word similarity simply as the amount of character or phoneme insertions, deletions, and substitutions required to make one word equivalent to the other. However, in language change, certain sound correspondences are more likely than others. Several methods rely on such linguistic knowledge by converting sounds into sound classes according to phonetic similarity [**?**]. For example, [15] consider a pair of words to be cognates when they match in their first two consonant classes.

In addition to such heuristics, a common approach to automatic cognate identification is to use edit distance metrics using weightings based on previously identified regular sound correspondences. Such correspondences can also be learned automatically by aligning the characters of a set of initial cognate pairs [3, 7]. In addition to sound correspondences, [14] and [6] also utilise semantic information of word pairs, as cognates tend to have similar, though not necessarily equivalent, meaning. Another method heavily reliant on prior linguistic knowledge is the LexStat method [9], requiring a sound correspondence matrix, and semantic alignment.

However, in the context of low-resource languages, prior linguistic knowledge such as initial cognate sets, semantic information, or phonetic transcriptions are rarely available. Therefore, cognate identification methods applicable to low-resource languages calls for unsupervised approaches. For example, [10] address this issue by investigating edit distance metrics based on embedding characters into a vector space, where character similarity depends on the set of characters they co-occur with. In addition, [12] investigate several unsupervised approaches such as hidden Markov models and pointwise mutual information, while also combining these with heuristic methods for improved performance.

3 Corpus

The initial plan is to base our method on the nominal XML dictionaries for the three Sami languages available on the Giellatekno infrastructure. Apart from just translations, these dictionaries contain also additional lexical information to a varying degree. The additional information which might benefit our research goals are cognate relationships, semantic tags, morphological information, derivation and example sentences.

For each noun the noun dictionaries, we produce a list of all its inflections in different grammatical numbers and cases. This is done by using a Python library called Uralic NLP [5], specialized in NLP for Uralic languages. Uralic NLP uses FSTs (finite-state-transducers) from the Giellatekno infrastructure to produce the different morphological forms.

We are also considering a possibility of including larger text corpora in these languages as a part of our method for finding cognates. However, theses languages

have notoriously small corpora available, which might render them insufficient for our purposes.

4 Future Work

Our research is currently at its early stages. The immediate future task is to start implementing different methods based on the previous research to solve the problem. We will first start with edit distance approaches to see what kind of information those can reveal and move towards a more complex solution from there.

A longer-term future plan is to include more languages into the research. We are also interested in a collaboration with linguists who could take a more qualitative look at the cognates found by our method. This will nourish interdisciplinary collaboration and exchange of ideas between scholars of different backgrounds.

We are also committed to releasing the results produced by our method to a wider audience to use and profit from. This will be done by including the results as a part of the XML dictionaries in the Giellatekno infrastructure and also by releasing them in an open-access MediaWiki based dictionary for Uralic languages [13] developed in the University of Helsinki.

References

1. Álgu-tietokanta. saamelaiskielten etymologinen tietokanta (Nov 2006), http://kaino.kotus.fi/algu/
2. Aikio, A.: The Saami loanwords in Finnish and Karelian. Ph.D. thesis, University of Oulu, Faculty of Humanities (2009)
3. Ciobanu, A.M., Dinu, L.P.: Automatic detection of cognates using orthographic alignment. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 99–105 (2014)
4. Häkkinen, K.: Suomen kirjakielen saamelaiset lainat. Teoksessa Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21, 161–182 (2007)
5. Hämäläinen, M.: UralicNLP (Jan 2018), https://doi.org/10.5281/zenodo.1143638, doi: 10.5281/zenodo.1143638
6. Hauer, B., Kondrak, G.: Clustering semantically equivalent words into cognate sets in multilingual lists. In: Proceedings of 5th international joint conference on natural language processing. pp. 865–873 (2011)
7. Kondrak, G.: Identification of cognates and recurrent sound correspondences in word lists. TAL 50(2), 201–235 (2009)
8. Koponen, E.: Lappische lehnwörter im finnischen und karelischen. Lapponica et Uralica. 100 Jahre finnisch-ugrischer Unterricht an der Universität Uppsala. Vorträge am Jubiläumssymposium 20.–23. April 1994 pp. 83–98 (1996)
9. List, J.M., Greenhill, S.J., Gray, R.D.: The potential of automatic word comparison for historical linguistics. PloS one 12(1), e0170046 (2017)
10. McCoy, R.T., Frank, R.: Phonologically informed edit distance algorithms for word alignment with low-resource languages. Proceedings of the Society for Computation in Linguistics (SCiL) 2018 pp. 102–112 (2018)

11. Moshagen, S.N., Pirinen, T.A., Trosterud, T.: Building an open-source development infrastructure for language technology projects. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16. pp. 343–352. No. 85, Linkping University Electronic Press; Linkpings universitet (2013)
12. Rama, T., Wahle, J., Sofroniev, P., Jäger, G.: Fast and unsupervised methods for multilingual cognate clustering. arXiv preprint arXiv:1702.04938 (2017)
13. Rueter, J., Hämäläinen, M.: Synchronized mediawiki based analyzer dictionary development. In: Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages. pp. 1–7 (2017)
14. St Arnaud, A., Beck, D., Kondrak, G.: Identifying cognate sets across dictionaries of related languages. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2519–2528 (2017)
15. Turchin, P., Peiros, I., Murray, G.M.: Analyzing genetic connections between languages by matching consonant classes. Vestnik RGGU. Seriya "Filologiya. Voprosy yazykovogo rodstva", (5 (48)) (2010)