digital humanities

Benelux

Antwerp

2015

DIXIT

University of Antwerp

DHu.f

DHBENELUX2

BOOK OF ABSTRACTS FOR THE
SECOND DIGITAL HUMANITIES BENELUX CONFERENCE

Antwerp, 8 & 9 June, 2015

This book of abstracts was typesetted in LaTeX by Mike Kestemont, using `ec2latex`, developed by Chris Emmery, and styled based on `arsclassica`. The `ec2latex` source is open and available on github.

Book of abstracts of the Second Digital Humanities Benelux Conference (DHBenelux2).

Antwerp, Arts Faculty

Edited by: DHBenelux Association (`www.dhbenelux.org`)
Cover design: Wout Dillen

PREFACE

The Digital Humanities Benelux Conference is an annual scholarly conference on all research related to the Digital Humanities (DH). This book of abstracts contains the contributions to the second edition of the conference, held on June 8 & 9 2015, in Antwerp, Belgium. The conference covers a broad range of discipline-specific and interdisciplinary research within the humanities. The aim of the conference is to build an inclusive DH community and foster research and collaboration. The first edition was held in The Hague, The Netherlands in 2014.

This year's conference features two keynote speakers, William Noel, Director of The Kislak Center for Special Collections, Rare Books and Manuscripts at the University of Pennsylvania; and Elena Pierazzo, Professor of Italian Studies and Digital Humanities at Stendhal University, Grenoble III. Similar to last year, the conference invited abstracts only and used a light review process where each abstract is reviewed by at least three reviewers and judged on clarity, relevance and novelty.

In total, 81 abstracts were submitted (compared to 62 in 2014) of which 76 were accepted (94% acceptance rate). The program consists of 57 oral presentations, 13 posters, 20 demos and 3 panels, with many abstracts being presented both as oral presentation and poster or demo. In terms of research directions, the popular themes of last year, e.g. Curation, Cyber Culture, Distant Reading, Networks and Topic Modeling, are well-represented again, with no significant shifts, but many other topics are covered as well. With such a diverse program, it seems the DHBenelux community is thriving!

We are very happy to be able to announce that this year we will be cooperating with the journal *Digital Humanities Quarterly* to publish a special issue based on the best papers selected from the body of work presented at the conference.

We thank all Program Committee members for their timely and constructive reviews and the local organizers for their enormous effort in organizing what will no doubt be a great conference with exciting events in a wonderful city. We gratefully acknowledge the support of EasyChair for organizing paper submission and reviewing.

We hope you enjoy the conference!

Amsterdam, June 8, 2015

Program chairs:
Joris van Zundert
Marijn Koolen

COMMITTEES

Organizing Committee

Elli Bleeker, University of Antwerp
Thomas Crombez, Royal Academy of Fine Arts & University of Antwerp
Walter Daelemans, University of Antwerp
Katrien Deroo, Ghent University
Wout Dillen, University of Antwerp
Ellen Janssens, University of Antwerp
Iason Jongepier, University of Antwerp
Aodhán Kelly, University of Antwerp
Mike Kestemont, University of Antwerp
Trudi Noordermeer, University of Antwerp
Saskia Scheltjens, Ghent University
Ben Verhoeven, University of Antwerp
Dirk Van Hulle, University of Antwerp

Program Committee

Joris J. van Zundert (Program Chair), Huygens KNAW
Marijn Koolen (Vice Program Chair), University of Amsterdam
Florentina Armaselu, CVCE Luxembourg
Paul Bertrand, Université Catholique de Louvain
Marnix Beyen, University of Antwerp
Rens Bod, University of Amsterdam
Barbara Bordalejo, KULeuven
Steven Claeyssens, Royal Library, The Hague
Sally Chambers, Ghent University
A. Seza Doğruöz, Tilburg University
Seth Van Hooland, Université Libre de Bruxelles
Catherine Emma Jones, CVCE Luxembourg
Folgert Karsdorp, Meertens Institute
Anne Roekens, Université de Namur
Els Stronks, Utrecht University
Karina van Dalen-Oskam, University of Amsterdam & Huygens KNAW
Antal van den Bosch, Radboud University Nijmegen
Nicoline van der Sijs, Radboud University Nijmegen
Christophe Verbruggen, Ghent University
Lars Wieneke, CVCE Luxembourg
Sally Wyatt, Maastricht University & KNAW

Sponsoring and support

The support of sponsors and exhibitors at DH Benelux 2015 has been crucial in allowing us to successfully host this event. We would like to offer our sincere thanks to Adam Matthew Digital, Taalmonsters, eHumanities KNAW and Ashgate Publishing. Special mention also goes to Gale Cengage both for sponsoring the event and providing funding towards the Early Career Bursaries. At the local level, we gratefully acknowledge the financial support of the departments of Literature, Linguistics and History at the Arts Faculty of the University of Antwerp, as well as the University Library and the department of 'Universiteit en Samenleving' at the University of Antwerp. Finally, we thank the research community 'DHuF: Digital Humanities Flanders' for its financial aid, sponsored via the Research Foundation of Flanders (www.fwo.be).

PAST, PRESENT AND FUTURE CONFERENCE MEETINGS

- June 12–13 2014 The Hague, The Netherlands
- June 8–9 2015 Antwerp, Belgium
- — 2016 Luxemburg

Gold level sponsor

Silver level sponsor

Bronze level sponsors

Other sponsors

GALE ARTEMIS
PRIMARY SOURCES

Unifies Gale Digital Collections into the one environment to enable innovative research for scholars and students

Now you can cross-search Gale Digital Collections all on the one platform including:

- Eighteenth Century Collections Online
- Nineteenth Century Collections Online
- Gale Historical Newspapers from Gale NewsVault
- The Making of the Modern World
- The Making of Modern Law

Contact your Gale rep today for more information about these remarkable primary source collections in Gale Artemis: Primary Sources or email **emea.galereply@cengage.com**

All libraries who own two or more Gale Digital Collections have automatic access to the Gale Artemis: Primary Sources experience. To find out more – email **emea.galereply@cengage.com**

GALE CENGAGE Learning

gale.cengage.co.uk

PROGRAMME

Monday 08 June	**Day 1**		
10:30-11:00	**Registration:** in the Dürer room		
11:00-12:00	**Plenary Session:** Keynote by Will Noel in the Tassis room		
Session A 12:00-13:00	**Reflection & Criticism I** *Room: Tassis* *Chair: S. Chambers*	**Cyber Culture I** *Room: Gresham* *Chair: P. Boot*	**Distant Readings I** *Room: Elsschot* *Chair: K. van Dalen-Oskam*
	Digital humanities: New toys for the boys? Sally Wyatt	**Monitoring Online User Behaviour. The case of the Newstracker** Martijn Kleppe, Irene Costera Meijer	**Writing Songs into Literary History with Digital Text Mining** Lisanne Vroomen
	The Digital Humanities cycle: hermeneutics, heuristics, and source criticism in a digital age Jesper Verhoef, Melvin Wevers	**On Seeking the Other: An Outlook on Digital Doppelgänger Trends** Alia Soliman	**HEEM, a Complex Model for Mining Emotions in Historical Text** Inger Leemans, Janneke M. van der Zwaan, Isa Maks, Erika Kuijpers
	Mapping Digital Humanities projects Stef Scagliola, Barbara Safradin, Almila Akdag, Sally Wyatt, Andrea Scharnhorst	**Ghost in the Machin...ima: Reflecting on Machinima's Ability of Going Mainstream** Aris Emmanouloudis	**God does not sing. Identification of participants in Psalm 75** Christiaan Erwich, Wido van Peursen
13:00-14:00	**Lunch:** in the Dürer room		

Session B 14:00-15:00	**Digital Scholarly Editing I** *Room: Tassis* *Chair: J. van Zundert*	**Linked Open Data** *Room: Gresham* *Chair: I. Hendrickx*	**Panel I** *Room: Elsschot* *Chair: C. Rasterhof*
	Specifying a system for collaborative online scholarly editing Peter Robinson	**Semantic Enrichment of a Multilingual Archive with Linked Open Data** Max de Wilde, Simon Hengchen	**Historical data exploration: Amsterdam's creative landscape, 1600-present** Leonor Álvarez Francés, Rosa Merino Claros, Harm Nijboer, Julia Noordegraaf, Claartje Rasterhoff
	Choosing publics as a scholarly act: challenges for creators of digital editions at the point of inception Elli Bleeker, Aodhán Kelly	**The possibilities and challenges of using linked data for academic research: the case of the Talk of Europe project** Laura Hollink, Martijn Kleppe, Max Kemman, Astrid van Aggelen, Willem Robert van Hage	
	Editions and editors of Greek papyrological texts Mark Depauw, Yanne Broux	**Globalising a Pan-European Common Names Webservice @ the Cloud: 5 steps towards Linked Open Data Humanities** Eveline Wandl-Vogt, Heimo Rainer, Thierry Declerck	
15:00-15:30	**Coffee break:** in the Dürer room		

Session C 15.30-16.30	**Panel II** *Room: Tassis* *Chair: D. van Hulle*	**Reflection & Criticism II** *Room: Gresham* *Chair: S. Wyatt*	**Digital Scholarly Editing II** *Room: Elsschot* *Chair: B. Bordalejo*	**Curation & Collection I** *Room: Prentenkabinet* *Chair: M. Beyen*
	Digitization and Exogenesis Wout Dillen, Dirk van Hulle, Tom de Keyser, Ronan Crowley, Vincent Neyt	**A method for cleaning 19th century text with examples from Transactions of the Royal Irish Academy 1800–1899** Emma Clarke, Alexander O'Connor	**Annotating Medieval Manuscript Layout** Hannah Busch, Jochen Graf, Philipp Vanscheidt	**Bridging Knowledge Collections: Integrating the museum and library systems at the Royal Museums of Art and History (RMAH), Brussels** Ellen van Keer
		User Required? On the Value of User Research in the Digital Humanities Max Kemman, Martijn Kleppe	**Europe's Beginnings through the Looking Glass: Publishing Historical Documents on the Web Using EVT** Roberto Rosselli Del Turco, Florentina Armaselu, Lars Wieneke, Chiara Di Pietro, Raffaele Masotti	**Measuring the Use of Collections Before and After Publication in Wikimedia** Trilce Navarrete
		Assessing the detection of text-reuse Peter Boot	**A Digital Scholarly Edition for Historians. The Case of Matthew of Edessa** Tara L. Andrews	**Location Extraction Tool** Rosa Merino Claros, Alex Olieman
17:45-19:00	**Poster reception:** in Antwerp's Historical Zoo (next to the Central Station)			
19:00-22:00	**Conference dinner, and nocturnal visit to the Zoo:** in Antwerp's Historical Zoo (next to the Central Station)			

Tuesday 09 June	**Day 2**		
Session D 09.00-10.20	**Distant Readings II** *Room: Tassis* *Chair: T. Andrews*	**Curation & Collection II** *Room: Gresham* *Chair: T. Noordermeer*	**Geospatial Applications** *Room: Elsschot* *Chair: W. Dillen*
	Beyond the Book: globalization in literature from a digital perspective Floor Buschenhenke, Karina van Dalen-Oskam, Carlos Martinez-Ortiz, Marijn Koolen	**Varieties in contemporary Dutch: Combining the research possibilities offered by MIMORE and GrETEL** Liesbeth Augustinus, Ineke Schuurman, Sjef Barbiers	**GIStorical Antwerp: a micro-level data tool for the study of past urban societies, test-case: Antwerp** Iason Jongepier, Ellen Janssens
	Literary constellations. A Digital Humanities approach to the study of literary salons in Mexico during the 19th century Silvia Gutiérrez	**How Can Language Technology Fight Against Language Death?** Ivett Benyeda, Eszter Simon, Péter Koczka	**MAGIS Brugge: a 16th-century bird's-eye view on Bruges as a digital stage for public urban history** Elien Vernackt, Bram Vannieuwenhuyze
	Visual Hermeneutics Peter Verhaar		**Analyzing the Dutch-Asiatic Trade in the 17th and 18th Centuries by Using a Spatial and Quantitative Approach** Erik van Zummeren, Robert-Jan Korteschiel
	The Document as Event. Applying automatic semantic role labeling to collections of theatre reviews Thomas Crombez		**Examining the Bomb Sight augmented reality app for exploring location based historical data** Catherine Emma Jones, Dan Karran, Patrick Weber
10:20–11:00	**Coffee break:** in the Dürer room		

Session E 11:00-12:00	**Networks & Topics I** *Room: Tassis* *Chair: A. van den Bosch*	**Curation & Collection III** *Room: Gresham* *Chair: E. Bleeker*	**Distant Readings III** *Room: Elsschot* *Chair: M. Koolen*
	Measuring Impact. Using Topic Modeling to analyse the influence of the Second Vatican Council on Dutch public debate Maarten van den Bos	**The History of Cultural Heritage Collections: Exploring New Approaches to Analysis and Visualization** Toby Burrows	**Thomas Kling's Bakchische Epiphanien – Projekt "Vorzeitbelebung" Reconstructing the Digital Writing Process from a Hard Drive in the Thomas Kling Archive (Digital Forensics)** Thorsten Ries
	TEI Annotation and Network Analysis of Diplomatic History Documents Florentina Armaselu, Marten Düring, Veronica Martins	**Event-based object identification** Alina Saenko	**Polemics Visualized: Morphological Analysis for Syriac** Hannes Vlaardingerbroek, Marieke van Erp, Wido van Peursen
	Reporting the Empire: The branding of Metropolises and Empire in the Pall Mall Gazette 1870–1900 Tessa Hauswedell, Melvin Wevers	**Enabling personal authoring and narrative making: Sustainable reuse of digital objects for interactive teaching and learning tools** Catherine Emma Jones, Lars Wieneke, Alexandre Genon, Frederic Reis, Frederic Allemand	**Combining quantitative and qualitative methods in digital analyses of literary style** J. Berenike Herrmann
12:00-13:00	**Lunch:** in the Dürer room		

Session F 13:00-14:20	**Networks & Topics II** *Room: Tassis* *Chair: L. Wieneke*	**Cyber Culture II** *Room: Gresham* *Chair: W. Daelemans*	**Curation & Collection IV** *Room: Elsschot* *Chair: T. Crombez*
	Automated historical judgement extraction: Analyzing perspectives on big and small heroes through NLP Miel Groten, Yassine Karimi, Serge ter Braake, Antske Fokkens	**Mapping urban multilingualism through Twitter** Enrique Manjavacas, Ben Verhoeven	**The TRAME Project – Text and Manuscript Transmission of the Middle Ages in Europe** Emiliano Degl'Innocenti, Alfredo Cosco
	Tourist or Pilgrim? Modeling two types of travel bloggers Tom van Nuenen, Suzanne van der Beek	**Automatically identifying periodic social events from Twitter** Florian Kunneman, Antal van den Bosch	**Visualizing the Dutch Folktale Database** Iwe Everhardus Christiaan Muiser, Mariet Theune
	What makes dream text dreamy? Antal van den Bosch, Iris Hendrickx, Maarten van Gompel, Ali Hürriyetoğlu, Folgert Karsdorp, et. al	**TheRiddlerBot: A Next Step on the Ladder Towards Computational Creativity** Ben Verhoeven, Iván Guerrero, Francesco Barbieri, Pedro Martins, Rafael Pérez y Pérez	**Visualizing medieval book production: Data visualization in medieval manuscript studies** Giulio Menna, Marjolein de Vos
	Modelling Discussion Topics to Improve News Article Tagging Chris Emmery, Menno van Zaanen	**Topics in Dutch Minority Languages on Twitter** Anna Katrine Jørgensen, Lysbeth Jongbloed-Faber, Jolie van Loo	**Using Parallel Data to Improve Part-of-speech Tagging of 17th Century Dutch** Dieuwke Hupkes, Rens Bod
14:20-16:00	**Demo session:** in the Dürer room		
16:00-16:30	**Plenary Session:** Keynote by Elena Pierazzo in the Tassis room		
16:30-17:00	**Closing panel:** in the Tassis room		

ABSTRACTS

Keynotes **24**

Oral Presentations **6**

A Digital Scholarly Edition for Historians. The Case of Matthew of Edessa 7
Tara L. Andrews

A method for cleaning 19th century text with examples from Transactions of the Royal Irish Academy 1800–1899 9
Emma Clarke, Alexander O'Connor

A repository for sources about late and post-colonial architecture and planning 12
Pauline K.M. van Roosmalen

Analyzing the Dutch-Asiatic Trade in the 17th and 18th Centuries by Using a Spatial and Quantitative Approach 14
Erik van Zummeren, Robert-Jan Korteschiel

Annotating Medieval Manuscript Layout 15
Hannah Busch, Jochen Graf, Philipp Vanscheidt

Assessing the detection of text-reuse 16
Peter Boot

Automated historical judgement extraction: Analyzing perspectives on big and small heroes through NLP 18
Miel Groten, Yassine Karimi, Serge ter Braake, Antske Fokkens

Automatically identifying periodic social events from Twitter 20
Florian Kunneman, Antal van den Bosch

Beyond the Book: globalization in literature from a digital perspective 21
Floor Buschenhenke, Karina van Dalen-Oskam, Carlos Martinez-Ortiz, Marijn Koolen

Choosing publics as a scholarly act: challenges for creators of digital editions at the point of inception 23
Elli Bleeker, Aodhán Kelly

Combining quantitative and qualitative methods in digital analyses of literary style . 24
J. Berenike Herrmann

Digital humanities: New toys for the boys? 26
Sally Wyatt

Digitization and Exogenesis 28
Wout Dillen, Dirk van Hulle, Tom de Keyser, Ronan Crowley, Tom de Keyser

Editions and editors of Greek papyrological texts 30
Mark Depauw, Yanne Broux

Enabling personal authoring and narrative making: Sustainable reuse of digital objects for interactive teaching and learning tools 32
Catherine Emma Jones, Lars Wieneke, Alexandre Genon, Frederic Reis, Alexandre Genon

Europe's Beginnings through the Looking Glass: Publishing Historical Documents on the Web Using EVT 35
Roberto Rosselli Del Turco, Florentina Armaselu, Lars Wieneke, Chiara Di Pietro, Lars Wieneke

Event-based object identification 38
Alina Saenko

Examining the Bomb Sight augmented reality app for exploring location based historical data 40
Catherine Emma Jones, Dan Karran, Patrick Weber

From lighthouse to the moon: a guiding light to the corpus of Jules Verne 42
Nicholas J. Hayward, George K. Thiruvathukal

GIStorical Antwerp: a micro-level data tool for the study of past urban societies, test-case: Antwerp 44
Iason Jongepier, Ellen Janssens

Ghost in the Machin...ima: Reflecting on Machinima's Ability of Going Mainstream . 45
Aris Emmanouloudis

Globalising a Pan-European Common Names Webservice @ the Cloud: 5 steps towards Linked Open Data Humanities 46
Eveline Wandl-Vogt, Heimo Rainer, Thierry Declerck

God does not sing. Identification of participants in Psalm 75 48
Christiaan Erwich, Wido van Peursen

HEEM, a Complex Model for Mining Emotions in Historical Text 49
Inger Leemans, Janneke M. van der Zwaan, Isa Maks, Erika Kuijpers

Historical data exploration: Amsterdam's creative landscape, 1600-present 51
Leonor Álvarez Francés, Rosa Merino Claros, Harm Nijboer, Julia Noordegraaf, Harm Nijboer

How Can Language Technology Fight Against Language Death? 53
Ivett Benyeda, Eszter Simon, Péter Koczka

Literary constellations. A Digital Humanities approach to the study of literary salons in Mexico during the 19th century 55
Silvia Gutiérrez

Location Extraction Tool 57
Rosa Merino Claros, Alex Olieman

MAGIS Brugge: a 16th-century bird's-eye view on Bruges as a digital stage for public urban history 58
Elien Vernackt, Bram Vannieuwenhuyze

Mapping Digital Humanities projects 60
Stef Scagliola, Barbara Safradin, Almila Akdag, Sally Wyatt, Almila Akdag

Mapping urban multilingualism through Twitter 62
Enrique Manjavacas, Ben Verhoeven

Measuring Impact. Using Topic Modeling to analyse the influence of the Second Vatican Council on Dutch public debate 63
Maarten van den Bos

Measuring the Use of Collections Before and After Publication in Wikimedia 65
Trilce Navarrete

Modelling Discussion Topics to Improve News Article Tagging 68
Chris Emmery, Menno van Zaanen

Monitoring Online User Behaviour. The case of the Newstracker 69
Martijn Kleppe, Irene Costera Meijer

On Seeking the Other: An Outlook on Digital Doppelgänger Trends 71
Alia Soliman

People on the Move and Cultural Heritage in a Digital Era. Participative and biographical collecting at the Red Star Line Museum Antwerp 73
Marie-Charlotte Le Bailly

Polemics Visualized: Morphological Analysis for Syriac 76

Hannes Vlaardingerbroek, Marieke van Erp, Wido van Peursen

Reporting the Empire: The branding of Metropolises and Empire in the Pall Mall Gazette 1870–1900 78
Tessa Hauswedell, Melvin Wevers

Semantic Enrichment of a Multilingual Archive with Linked Open Data 80
Max de Wilde, Simon Hengchen

Specifying a system for collaborative online scholarly editing 82
Peter Robinson

TEI Annotation and Network Analysis of Diplomatic History Documents 85
Florentina Armaselu, Marten Düring, Veronica Martins

The Digital Humanities cycle: hermeneutics, heuristics, and source criticism in a digital age 87
Jesper Verhoef, Melvin Wevers

The Document as Event. Applying automatic semantic role labeling to collections of theatre reviews 89
Thomas Crombez

The History of Cultural Heritage Collections: Exploring New Approaches to Analysis and Visualization 90
Toby Burrows

The TRAME Project – Text and Manuscript Transmission of the Middle Ages in Europe 92
Emiliano Degl'Innocenti, Alfredo Cosco

The curation of sound archives: the Dutch Dialect Database 94
Douwe Zeldenrust

The possibilities and challenges of using linked data for academic research: the case of the Talk of Europe project 96
Laura Hollink, Martijn Kleppe, Max Kemman, Astrid van Aggelen, Max Kemman

TheRiddlerBot: A Next Step on the Ladder Towards Computational Creativity 98
Ben Verhoeven, Iván Guerrero, Francesco Barbieri, Pedro Martins, Francesco Barbieri

Thomas Kling's Bakchische Epiphanien – Projekt "Vorzeitbelebung" Reconstructing the Digital Writing Process from a Hard Drive in the Thomas Kling Archive (Digital Forensics) 99
Thorsten Ries

Topics in Dutch Minority Languages on Twitter 101
Anna Katrine Jørgensen, Lysbeth Jongbloed-Faber, Jolie van Loo

Tourist or Pilgrim? Modeling two types of travel bloggers 103
Tom van Nuenen, Suzanne van der Beek

User Required? On the Value of User Research in the Digital Humanities 104
Max Kemman, Martijn Kleppe

Using Parallel Data to Improve Part-of-speech Tagging of 17th Century Dutch 106
Dieuwke Hupkes, Rens Bod

Varieties in contemporary Dutch: Combining the research possibilities offered by MIMORE and GrETEL 107
Liesbeth Augustinus, Ineke Schuurman, Sjef Barbiers

Visual Hermeneutics 109
Peter Verhaar

Visualizing medieval book production: Data visualization in medieval manuscript studies 110
Giulio Menna, Marjolein de Vos

Visualizing the Dutch Folktale Database 111
Iwe Everhardus Christiaan Muiser, Mariet Theune

Visualizing the Narratives of European Integration 113
Laurie de Zwart, Sarah Döking, Nicky van Rijsbergen, Tijmen Weber,
Nicky van Rijsbergen, Zsófi Bognár, Adrienn Adolf, Réka Köcsky, Zíta Huszthy, Renata Hrecska, Iris Hendrickx, Antal van den Bosch, Zsolt Almási

What makes dream text dreamy? 115
Antal van den Bosch, Iris Hendrickx, Maarten van Gompel, Ali Hürriyetoğlu,
Maarten van Gompel, Florian Kunneman, Louis Onrust, Martin Reynaert, Wessel Stoop

Writing Songs into Literary History with Digital Text Mining 116
Lisanne Vroomen

'Catalogue these books': Digital Editions and the Digital Library 119
Ronan Crowley

Presented as demos only **122**

A Lexicon of Scholarly Editing 123
Wout Dillen

A Preview of CENDARI. A Digital Research Infrastructure 124
Stijn van Rossem

Bridging Knowledge Collections: Integrating the museum and library systems at the Royal Museums of Art and History (RMAH), Brussels 125
Ellen van Keer

Ebacs: a Minimalistic Conference Manager 128
Chris Emmery

Iter Community: Enabling Social Bibliography and User Project Creation 129
Shawn DeWolfe, Daniel Sondheim, Matthew Hiebert, William Bowen, Matthew Hiebert

Nederlab, online laboratory for humanities research on Dutch text collections 131
Hennie Brugman

The Beckett Digital Manuscript Project and Beckett's Personal Library 132
Dirk van Hulle, Vincent Neyt

Presented as posters only **134**

DARIAH and the Benelux 135
Sally Chambers, Maarten Hoogerwerf, Jan van der West, Marianne Backes

DiXiT – Digital Scholarly Editions Initial Training Network 137
Elli Bleeker, Aodhán Kelly

Digital Scholarship at the Koninklijke Bibliotheek 138
Lotte Wilms, Steven Claeyssens

KEYNOTES

Keynote Speakers

William Noel

Our first keynote speaker is Will Noel, the current Director of The Kislak Center for Special Collections, Rare Books and Manuscripts and The Schoenberg Institute for Manuscript Studies at the University of Pennsylvania. Before that, Will was Curator of Manuscripts and Rare Books at The Walters Art Museum, Baltimore, Maryland and British Academy Post-Doctoral Research Fellow in the Department of History of Art, Cambridge University. Even before that he was Assistant Curator of Manuscripts at The J. Paul Getty Museum in Los Angeles. Will's area of expertise are illuminated manuscripts, a field in which he has directed various projects. He is probably best known for his work on the Archimedes Palimpsest project, as well as various digitization initiatives at the Walters, and more recently in Philadelphia, such as the acclaimed OPenn Digital Resources Online Platform. A delightful TED-talk of his on the Archimedes codex can be viewed at: `http://www.ted.com/talks/william_noel_revealing_the_lost_codex_of_archimedes`. Will is a fierce advocate of open data ("Data is going to die if it's not used") and is committed to creating machine-readable data sets that can be used for free, by anybody, for any purpose. In 2013, Will was honored as a White House Champion of Change for his visionary commitment to open science. He maintains a lively Twitter feed under the name `@willnoel`.

Elena Pierazzo

Elena Pierazzo is a world-renown protagonist in Digital Humanities and digital scholarly editing in particular. Currently Elena is Professor of Italian Studies and Digital Humanities at the University of Grenoble 3 'Stendhal'; formerly she was Lecturer at the Department of Digital Humanities at King's College London where she was the coordinator of the MA in Digital Humanities. She has a PhD in Italian Philology: her specialism is Italian Renaissance texts, digital edition of Early Modern and modern draft manuscripts, and text encoding. She has published and presented papers at international conferences in Renaissance liter-

ature, digital editions, text encoding theory and Italian linguistics. She is the Chair and C.E.O of the Text Encoding Initiative and involved in the TEI user-community, with a special interest in the transcription of modern and medieval manuscripts. She co-chairs the working group on digital editions of the European Network NeDiMAH. Several scholars in the community are very much looking forward to her forthcoming monograph called *Digital Scholarly Editing. Theories, Models, and Methods*, which will be published with Ashgate in August. Her devotion to the TEI becomes all the more clear through the name of her cat 'Tag'. Elena too is very active on Twitter (@epierazzo).

ORAL PRESENTATIONS

A Digital Scholarly Edition for Historians. The Case of Matthew of Edessa

Tara L. Andrews
University of Bern
tara.andrews@kps.unibe.ch

The Chronicle of Matthew of Edessa is well-known to most scholars of the medieval Near East and the First Crusade, both for the wealth of information that it includes and for the apparent ignorance and naïveté of its author, an Armenian priest living in the Crusader county of Edessa and writing between 1110 and 1132. Although the text is widely used by historians, it has never yet received a critical edition, and has not been published in any edition at all since 1898 (Uŕhayec'i, 1898). Surviving in some 35 manuscripts, all of which were copied at least 450 years after the death of the author, the Chronicle is now the subject of a fully digital text edition currently in preparation.

The challenge that the Chronicle presents is that it is not only a text that deserves a full philological treatment, but that it must be further annotated and presented as a historical work – as a platform for the study of a time and a place that many historians wish to know about but that is familiar to very few. This means that the work on the edited text encompasses not only the state-of-the-art methods familiar to digital philologists – full transcription of manuscripts, palaeographical markup using the vocabulary of the Text Encoding Initiative (TEI Consortium, 2015), automatic manuscript collation with CollateX (Dekker, Hulle, Middell, Neyt, and Zundert, 2014), stemmatic analysis through the tools provided by Stemmaweb (Andrews and Macé, 2013), and publication of the complete set of transcriptions along with an editorial reconstruction of the text – but that the commentary on the text must also take advantage of the digital form and the online resources that exist for a historical study of the period. Place names are not simply tagged; their probable locations are resolved, and they are displayed geographically insofar as possible. Personal names and ethnographic labels are not merely indexed; they are linked, wherever possible, to prosopographical databases or even relevant pages on Wikipedia. The annalistic nature of the Chronicle then allows a scholar to trace the movement of individuals and groups through time and space, making the textual edition itself a platform for the study of medieval history.

The presentation portion of this submission will cover the working methods of the project as they have been developed over the last four years, its technological underpinnings and software architecture, and the lessons that have been learned along the way. The submission also includes a demonstration of the edition-in-progress, so that interested scholars – particular others working on historical texts of a similar nature – may evaluate our work for their own needs.

References

Andrews, T. L., and Macé, C. (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing*, 28(4): 504–21. 10.1093/llc/fqto32.

Dekker, R. H., Hulle, D. van, Middell, G., Neyt, V., and Zundert, J. van. (2014). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Literary and Linguistic Computing*, fqu007.10.1093/llc/fqu007.
Matthew of Edessa (Matt'ēos Uṙhayec'i). (1898). Žamanakagrut'iwn. Vałaršapat.
TEI Consortium (Ed.). (2015, April 6). *Guidelines for Electronic Text Encoding and Interchange*. Version 2.8.0 (`http://www.tei-c.org/p5/`).

A method for cleaning 19th century text with examples from Transactions of the Royal Irish Academy 1800–1899

Emma Clarke
Trinity College Dublin
Clarkee8@tcd.ie

Alexander O'Connor
Trinity College Dublin
Alex.Oconnor@scss.tcd.ie

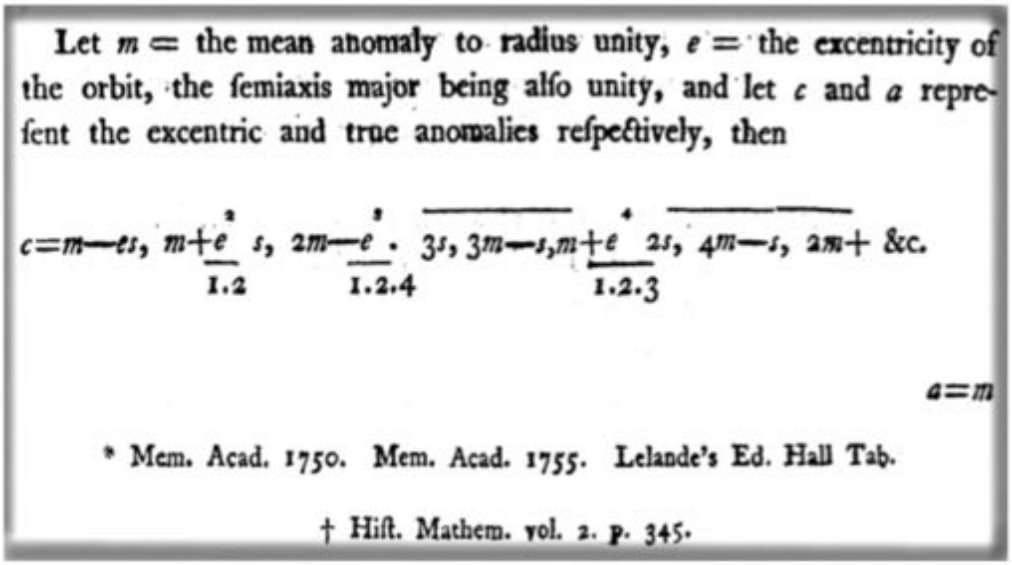
This paper presents the cleaning pipeline, which was built on a corpus of nineteenth century scholarly articles – Transactions of the Royal Irish Academy [RIA] 1800-1899; and discusses the effect of different attempts to regularise and normalise the machine mediation of the insight. One of the specific challenges which arises when attempting to carry out machine assisted textual analysis on a corpus such as *Transactions* is how to extract stable, insightful patterns from data which is difficult for the machine to understand. The articles within the corpus were all published between 1800 and 1899 and were digitised using OCR techniques. The digitisation was carried out as part of an unrelated project, a common situation for DH researchers, which limits control over quality and suitability.

This corpus presented characteristic challenges incorporated in the articles. The following factors had to be taken into consideration when cleaning the *Transactions* corpus (see the first figure below for specific examples from Transactions of the RIA (Brinkley, 1803)):

- the inclusion of technical and structural content (tables, references, symbols, equations)
- the use of multilingual and specialised language
- the importance of multi-word expressions (Sag et al., 2002)
- polysemy (words or phrases which have many different meanings)
- homonymy (words which are spelled the same, but have different meanings)
- other linguistic variations (abbreviations: Vol. for volume, Fig. for figure etc.)

Cleaning process and collection of resources

One of the key challenges in any automatic approach is the need to address the inconsistency, complexity and general noise which exists in real data. This is especially true in the case of a corpus such as this, as the content is over a wide diachronic range (affecting the typography amongst other things), has high subject-matter heterogeneity, and was converted using OCR techniques. A key aspect of this work was that there was no access to the underlying documents, to higher-quality OCR, or to reference texts. This meant that many of the approaches to performance improvement were not available. Previous approaches to cleaning 19th century OCR text include Underwood's normalisation of OCR errors (Underwood, 2013) and Jockers's combination of an expanded stop word list and part of speech tagging (Jockers, 2013). However, due



Let m = the mean anomaly to radius unity, e = the excentricity of the orbit, the ſemiaxis major being alſo unity, and let c and a repreſent the excentric and true anomalies reſpectively, then

$$c = m - es, \; m + \frac{e^2}{1.2}s, \; 2m - \frac{e^3}{1.2.4}\cdot 3s, 3m - s, m + \frac{e^4}{1.2.3}2s, \; 4m - s, \; 2m + \text{ \&c.}$$

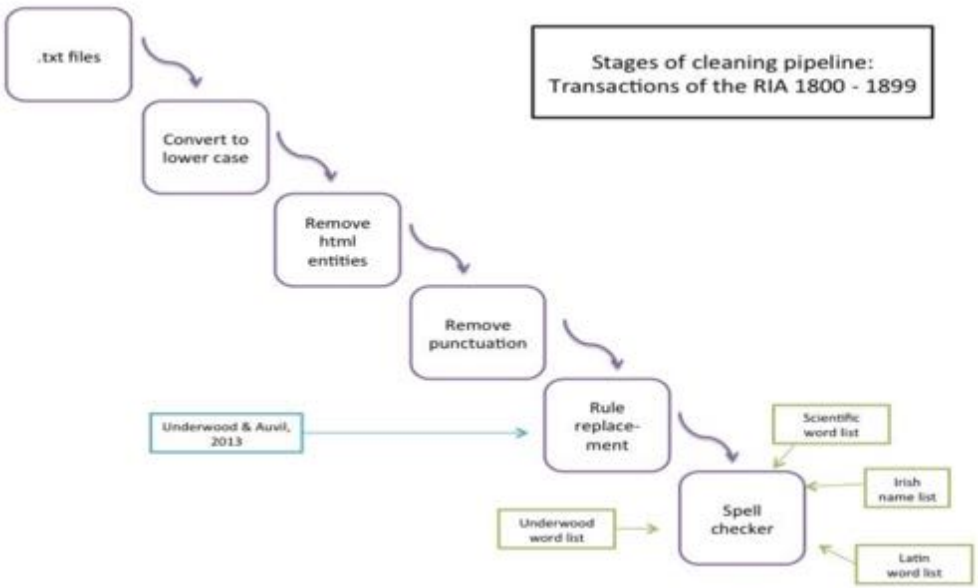
$a = m$

* Mem. Acad. 1750. Mem. Acad. 1755. Lelande's Ed. Hall Tab.

† Hiſt. Mathem. vol. 2. p. 345.

Figure 3: Examples of challenging factors in *Transactions of RIA*.

to the mixed and technical nature of the corpus content and the fact that we needed to remove structural content such as tables, charts and figures, we chose the approach outlined in the flowchart below (Underwood and Auvil n.d.; Underwood 2012; Petrie n.d.).



Python scripts were developed to carry out each stage of the cleaning pipeline and these are available in the following Github repository: https://github.com/emmaclarke/TransactionsRIA_1800-99.

Conclusion

The paper provides insight into an increasingly common challenge facing digital researchers: the artefacts which they wish to investigate are complex, incorporating significant variation in language, formatting, typesetting and structure. Moreover, the artefacts under investigation are the 'born-digital' incarnations: no access is available to the originals, and the collection is small and unusual. The approach that we present is therefore relevant for many studies which seek to gain insight into the implicit meaning in document corpora, where complete control is not possible. The positive improvements found with this approach, along with discussions on how to customise, and more formally evaluate, for similar work in future, are intended to show the underlying value of the work beyond the conclusions drawn about these texts.

References

Brinkley, J., 1803. An Examination of Various Solutions of Kepler's Problem, and a Short practical Solution of That Problem Pointed out. *Trans. R. Ir. Acad.* 83–131.

Jockers, M., 2013. "Secret" Recipe for Topic Modeling Themes [WWW Document]. Matthew Jockers (http://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/) (accessed 8.26.13).

Petrie, J., n.d. Scientific word list for spell-checkers/spelling dictionaries [WWW Document]. *John Petrie's LifeBlag* (http://www.jpetrie.net/scientific-word-list-for-spell-checkersspelling-dictionaries/) (accessed 8.18.13).

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., 2002. Multiword Expressions: A Pain in the Neck for NLP, in: Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1–15.

Underwood, T., Auvil, L., n.d. Basic OCR correction [WWW Document]. *Uses Scale Lit. Study* (http://usesofscale.com/gritty-details/basic-ocr-correction/) (accessed 6.25.13).

Underwood, Ted, 2013. A half-decent OCR normalizer for English texts after 1700. *Stone Shell* (http://tedunderwood.com/2013/12/10/a-half-decent-ocr-normalizer-for-english-texts-after-1700/) (accessed 6.25.13)

Underwood, Ted, 2012. Flowchart for probabilistic OCR correction. *Uses Scale Lit. Study* (http://usesofscale.com/2012/10/14/probabilisticocr/) (accessed 6.25.13)

A repository for sources about late and post-colonial architecture and planning

Pauline K.M. van Roosmalen
TU Delft
P.K.M.vanRoosmalen@TUDelft.nl

In 2011, the Chair of History of Architecture and Urban Planning of Delft University of Technology (TU Delft) in the Netherlands, initiated the development of a repository for sources about colonial architecture and town planning (circa 1850–1970) in former European colonies. The motivation driving the creation of the repository was the recognition that although colonial (built) heritage has become a research topic of increasing importance, studies about the development and significance of late-colonial architecture and planning remain an under-researched topic. As this vacuum to a large degree can be contributed to restricted access to relevant sources, TU Delft embarked on the creation of a dedicated repository that offers online open access to geographically often dispersed and isolated sources: texts documents, photographs, films, maps and archives.

The repository's data model is designed and developed in close collaboration with collections in the Netherlands and the repository's targeted user commmunities. Taking their suggestions on board, TU Delft Library developed an relational data model that suggests and presents relations and links within the repository's content/objects. The repository thus enables its users to detect, unravel, visualise and understand connections that so far often have gone unnoticed and potentially will call for modifications and corrections of prevailing notions, interpretations and views on European colonial architecture and planning. TU Delft is currently expanding its collaboration with other Dutch and non-Dutch institutes.

Figure 4: Homepage `colonialarchitecture.eu`.

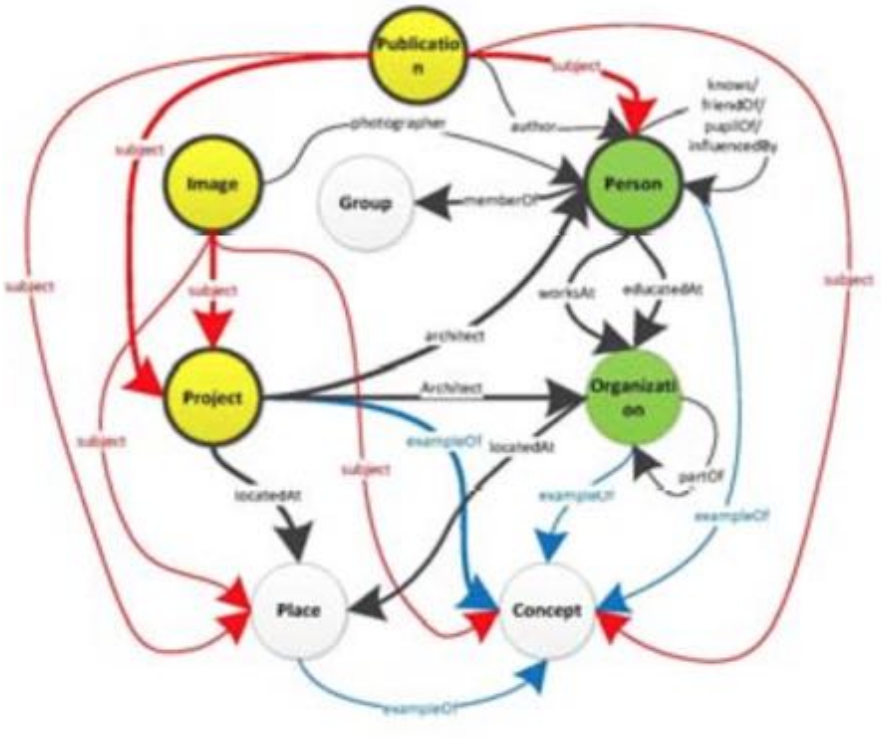


Figure 5: TU Delft repository data model ©.

Analyzing the Dutch-Asiatic Trade in the 17th and 18th Centuries by Using a Spatial and Quantitative Approach

Erik van Zummeren
University of Amsterdam
erikvanzummeren@gmail.com

Robert-Jan Korteschiel
University of Amsterdam
rjkorteschiel@gmail.com

In recent years the Dutch Huygens Institute has released several datasets on Dutch shipping in Asia. Examples include the digitisation of the Dutch-Asiatic Shipping dataset (DAS) and the Bookkeeper General Batavia database (BGB). These databases contain an impressive amount of information. It does not only contain basic information such as a place and date of departure of every single voyage that went from the Dutch Republic to Asia and vice versa. But it also contains, in the case of the BGB database, a detailed list of onboard commodities. However, in its present day usage these sets are mostly used as a work of reference. By visiting the DAS or BGB website you can get a good understanding of a single voyage, but it's quite difficult to see the bigger picture.

By creating a visualisation based on an open source GIS system we hope to tackle this problem. Our approach consists of plotting 18,000 voyages on a world map. Each of these voyages is linked with it's date of departure and arrival. The canvas of our application only shows the voyages from a user specified time period. For instance if a user selects March 1643 only those voyages taking place in March 1643 show up. Further insights can be gained by colourcoding voyages that contain certain commodities such as tin, opium or Chinese paper. The sum of their financial value is distributed in years and are plotted in a graph underneath the map. Events are also added to the map. They show how the voyages relate to certain events, and they provide a narrative.

Annotating Medieval Manuscript Layout

Hannah Busch
Trier Center for Digital Humanities
buschh@uni-trier.de

Jochen Graf
University of Cologne
jochen.graf@uni-koeln.de

Philipp Vanscheidt
Darmstadt University
of Technology
vanscheidt@linglit.tu-darmstadt.de

In addition to traditional human research with digitized historical sources, computer-assisted possibilities of research evolve: while digitized medieval manuscripts can be handled with programs and tools for supporting critical editing, it is also feasible to analyze the images themselves with trained algorithms and establish classifications for detecting codicological information on larger corpora and to evaluate them statistically. New and in some cases more precise information can be added to the catalogue data. Especially concerning measurements algorithmic results might be more accurate and reproducible. The quantitative evaluation can help to find hidden relations between layout features, genre and historical developments in a corpus or between different corpora. Both ways are used in the project "eCodicology" (www.ecodicology.org).

Using digitized images and catalogue information of the "Virtuelles Skriptorium St. Matthias" (stmatthias.uni-trier.de), 470 medieval books from many different areas of human knowledge and culture like theology, philosophy, or literature, and from the 8th to the 18th century are used for analyzing the practice and the history of this library of a Benedictine Abbey near the city of Trier. "eCodicology" searches for patterns in the development of the medieval book, different genres and historical ruptures insofar as they became manifest in layout features and relations between parts of the page like text and image areas or margins.

On the one hand, in cooperation with the virtual research environment "TextGrid" (www.textgrid.de) the opportunity is given to use the digital facsimiles of 470 codices in critical editions. On the other hand, by using the Software Workflow for the Automatic Tagging of (Medieval Manuscript) Images (SWATI) mise-en-page elements can be discovered automatically. These two ways – the human researcher's and the computer's way – cannot be strictly separated. The results can be ambiguous from a codicological point of view. In cases of unstable features like the position of a rubric on the page or in complex cases like multiple glossed texts, a human annotation can be useful. The software "SemToNotes" allows to select, delete and edit automatically recognized areas of the digital image and to annotate them by using some kind of codicological terminology. Thus, it becomes possible to adapt automatic results to research questions and use these results for the reconstruction of the complex relations between visual phenomena and semantic relevance in the history of the medieval book.

Assessing the detection of text-reuse

Peter Boot
Huygens ING
peter.boot@huygens.knaw.nl

One of the central questions in the history of literature and the history of ideas is always: to which other texts is the text that I am studying indebted? And which other texts has it influenced? The detection of quotation is a problem that has been addressed in a number of digital humanities projects, among others projects focusing on Latin poetry (Coffee et al., 2012), medieval encyclopedias (Mews et al., 2010), 18th-century French philosophy (Olsen et al., 2011), 19th-century French literature (Ganascia et al., 2014), and early US newspapers (Smith et al., 2013). A variety of tools has been developed, applying a variety of algorithms. Some of these tools started their life as tools for the detection of plagiarism (Kane and Tompa, 2011).

Most tools are based on the detection of matching word n-grams, and can be fine-tuned using a number of parameters, for instance allowing non-matching words between the matching words, by applying lemmatization, or by removing stop-words. Different parameter settings may be required for different languages (based on e.g. the amount of flexion in the language) and text collections (dependent on e.g. consistency of spelling, text quality (OCR vs. transcribed text) and genre).

The problem that this paper and demo will address is how to assess the suitability of a certain tool and parameter setting for specific textual corpora. Running a tool that looks for textual correspondences between two textual corpora results in a list of corresponding text places. The hits are true positives, false positives, or something in between. True correspondences that a tool does not detect remain invisible. Re-running the tool with different parameter settings results in an overlapping set of hits, and as there is a large number of different parameter settings, the researcher easily gets lost. Since the true amount of quotation is unknown and subject to interpretation, standard information retrieval measures to detect the best settings do not apply.

I my paper, I argue that what we need is (i) a database that stores all relevant information for each run of the detection tool: the tool version, the corpora properties, the applied settings and the results, (ii) an application that displays parameters and results and allows flexible navigation through the results, ad (iii) the facility to annotate the database at all levels: the program run, the processed texts, and the detected correspondences. The demo will show a prototype for such a tool. I developed the prototype to help me investigate intertextuality in the 17th-century emblem, a genre in which intertextuality played an important role. I used `Text::Pair` (`https://code.google.com/p/text-pair/`) to find textual relations between the emblem corpus digitized in the Emblem Project Utrecht (`http://emblems.let.uu.nl/`) and a number of other text corpora. The demo will show how the prototype helps assess the suitability of different parameter settings when using `Text::Pair` on these corpora.

References

Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R. & Jacobson, S. L. 2012. The Tesserae Project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28, 221-228.
Ganascia, J.-G., Glaudes, P. & Del Lungo, A. 2014. Automatic Detection of Reuses and Citations in Literary Texts. *Literary and Linguistic Computing*, 29, 412-421.
Kane, A. & Tompa, F. W. 2011. Janus: the intertextuality search engine for the electronic Manipulus florum project. *Literary and Linguistic Computing*, 26, 407-415.
Mews, C. J., Zahora, T., Nikulin, D. & Squire, D. 2010. The Speculum morale (c. 1300) and the study of textual transformations: a research project in progress. *Vincent of Beauvais Newsletter*, 35, 5-15.
Olsen, M., Horton, R. & Roe, G. 2011. Something borrowed: sequence alignment and the identification of similar passages in large text collections. *Digital Studies/Le champ numérique*, 2.
Smith, D. A., Cordell, R. & Dillon, E. M. Infectious texts: Modeling text reuse in nineteenth-century newspapers. *IEEE International Conference on Big Data, 2013*. IEEE, 86-94.

Automated historical judgement extraction: Analyzing perspectives on big and small heroes through NLP

Miel Groten
VU University Amsterdam
mielgroten@live.nl

Yassine Karimi
VU University Amsterdam
m.y.karimi@student.vu.nl

Serge ter Braake
VU University Amsterdam
s.ter.braake@vu.nl

Antske Fokkens
VU University Amsterdam
antske.fokkens@vu.nl

This year, a movie about Michiel de Ruyter (1607-1676), the most famous admiral from the Netherlands [1], was released. The fame of De Ruyter has inspired people for centuries. De Ruyter was even voted number seven in the 2004 Dutch tv elections for the grandest' Dutch person from history [2]. A search in the Google Ngram viewer shows that he also became increasingly known in the twentieth century in English language books [3]. Not everyone, however, believes he should be venerated as a hero. One group in particular, 'Michiel de Rover', protests against the movie and sees De Ruyter as a sea robber who played a doubtful role in Dutch slavery practices (Jensen 2015, Historiek 2015; Elsevier 2010 for earlier controversy).

De Ruyter is a good example, and many more can easily be found, of how the perspective on people changes over time and is dependent on social, cultural and political factors. It is difficult however, to see patterns in how time tells a different story. Such patterns are extremely interesting, because they could reveal a lot about historiography and identity. What attributes do 'heroes', or villains, get from historians over time? Is there a difference in the way women/men, politicians/painters, protestants/catholics are described over time? What topoi do historians use? Are there any 'national' characteristics discernible in historiographical sources? To answers such questions one would have to 'close read' an incredible number of sources from the past. Much more than any person could do in a reasonable amount of time. In this paper we apply NLP (automatic text analysis) tools to have a computer aid us in this task. This paper introduces our methodology and first results in using an NLP pipeline for automated historical judgement extraction.

For our purpose we use two extensive biographical dictionaries from the nineteenth and twentieth centuries to start with (Van der Aa 1852-1878, Blok and Molhuysen 1911-1937). They contain circa 45,000 short biographies on famous Dutch people from all history. There is a lot of overlap in the people described, which makes comparisons possible. Our approach in using NLP to mine perspectives from these texts is twofold. First, we analyze a relatively small selection of texts manually on positive and negative character traits which were attributed to people's characters and their deeds by the authors of the texts and by third parties mentioned in the texts. Second, we use an NLP pipeline that aims to identify positive and negative traits automatically. It is based on an existing pipeline that was originally built for event extraction, including a Named Entity Recognizer [4]. In this paper, we will introduce our methodology in

more detail, present the results (precision and recall) from our pipeline and will discuss what (preliminary) answers we were able to give to the kind of questions asked in the second paragraph.

Notes

[1] See http://www.admiralthemovie.com/.
[2] See http://en.wikipedia.org/wiki/De_Grootste_Nederlander.
[3] Google Books Ngram Viewer: https://books.google.com/ngrams/graph?content=Michiel+de+Ruyter&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2CMichiel%20de%20Ruyter%3B%2Cc0.
[4] For a schematic rendering of the tools used in the pipeline, see: http://www.biographynet.nl/wp-content/uploads/2014/12/Meertens.pdf.

References

Aa, A.J. van der (ed.) (1852-1878) *Het Biographisch Woordenboek der Nederlanden*, Haarlem: J.J. van Brederode.
Blok, P.J. and Molhuysen. P.C., (ed.) (1911-1937) *Nieuw Nederlandsch Biografisch Woordenboek*, Leiden: A.W. Sijthoff's Uitgeverij.
'Historiek'. (29 January 2015) Michiel de Rover. [Online] Available from: http://historiek.net/michiel-de-rover/47743/.
Jensen, Lotte (2015) 'Michiel de Ruyter kan dienen voor elk doel'. *De Volkskrant*, 30 January. p.20.
Elsevier. (19 June 2010) 'Stichting Michiel de Ruyter boos op 'zeeheld' Wilders'. [Online] Available from: http://www.elsevier.nl/Cultuur-Televisie/nieuws/2010/6/Stichting-Michiel-de-Ruyter-boos-op-zeeheld-Wilders-ELSEVIER268640W/.

Automatically identifying periodic social events from Twitter

Florian Kunneman
Radboud University
f.kunneman@let.ru.nl

Antal van den Bosch
Radboud University
a.vandenbosch@let.ru.nl

Many events that are referred to on Twitter are of a periodic nature, characterized by roughly similar intervals in between. Examples are biannual conferences, annual music festivals, weekly television programs and the full moon cycle. We propose a system that can automatically identify such events in an open-domain fashion from a long period of tweets. Based on such a system, any event in a periodic sequence can be enriched by linking it to previous and later editions. Furthermore, periodic event patterns can be leveraged to predict future events before they are mentioned on Twitter.

Our system is applied in three stages. Starting from tweets that go back to december 2010, it firstly extracts explicit event references to build a calendar of events. Secondly, the system finds similar events in time from these calendar entries and scores sequences by their periodicity. In the third stage, future events are predicted based on the most periodic sequences from the past.

We provide an overview of the methods and performance in these three stages, and show how the system is applied in an online setting. This work not only has relevance in a social media context. Many text collections carry periodic patterns, which can be leveraged to disclose knowledge about their domain.

Beyond the Book: globalization in literature from a digital perspective

Floor Buschenhenke
Huygens ING

Karina van Dalen-Oskam
Huygens ING
karina.van.dalen@huygens.knaw.nl

Carlos Martinez-Ortiz
Netherlands eScience Centre
c.martinez@esciencecenter.nl

Marijn Koolen
University of Amsterdam
marijn.koolen@uva.nl

Globalization is an important research topic in many fields, also in literary studies. In our project Beyond the Book, funded by the Netherlands eScience Center, we try to find a way to measure how "international" a novel is. The idea behind our project is that the topic and content of a novel from one country may have an appeal to readers in other countries if the novel, for instance, reflects a certain amount of cultural knowledge that is also available to readers in the other countries. It could also be the case that readers from other countries do not look for shared cultural elements, but prefer to read about exotic other cultures. This will depend on "fashions" in literature.

To establish the level of shared cultural knowledge or exoticness we use Wikipedia as a reference frame. Wikipedia is currently widely used in the analysis of cultural diversity. For example, Warncke-Wang et al. (2012) describe a way to compare different-language Wikipedias based on their size, amount of shared entries, and amount of unique entries per language. They also pay attention to the flow of translations between the Wikipedias. This results in an intriguing overview of content-related differences between Wikipedias. Eom & Shepelyansky (2013) provide rankings of persons based on their entries in different-language Wikipedias. Pfeil et al. (2006) analyse cultural differences based on measurable practices in collaborative authoring in Wikipedia. Their findings include some correlations with four cultural influences as proposed by Hofstede (1991). Comparable work on all five of Hofstede's dimensions of cultural differences (power distance, collectivism versus individualism, femininity versus masculinity, uncertainty avoidance, and the later added long-term versus short-term orientation) has been done by Hara et al. (2010).

Our own approach consists of a more detailed way to look at shared and unique knowledge. For this presentation we focus on the analysis of Wikipedia entries about proper names (named entities) that are found in novels. In particular, we are looking at the contributions to English Wikipedia entries from different countries as our source data. If entries dealing with topics show relatively more edits from users from certain countries compared to users from other countries, we interpret this as a sign of a cultural preference for the topic in those countries. The different cultural preferences as gathered from Wikipedia are then compared to a corpus of Dutch novels. In our presentation we will describe the pipeline of tools that we developed, show some first results, and present our ideas on next steps that could lead to more knowledge about the impact of globalization in literature.

References

Eom, Y., Shepelyansky, D. L. (2013) 'Highlighting entanglement of culture via ranking of multilingual Wikipedia articles', *PLoS ONE* 8(10): e74554. doi:10.1371/journal.pone.0074554

Hara, N., Shachaf, P., Hew, K. F. (2010) 'Cross cultural analysis of the Wikipedia community', *Journal of the American Society of Information Science and Technology* 61 (10), pp. 2097-2108.

Hofstede, G. H. (1991) *Cultures and organizations: Software of the mind*, London: McGraw-Hill.

Pfeil, U., Zaphiris, P., and Ang, C. S. (2006) 'Cultural differences in collaborative authoring of Wikipedia', *Journal of computer-mediated communication*, vol. 12, no. 1 [Online]. Available: `http://jcmc.indiana.edu/vol12/issue1/pfeil.html`.

Warncke-Wang, M., Uduwage, A., Dong, Z., and Riedl, J. (2012) 'In search of the ur-wikipedia: Universality, similarity, and translation in the wikipedia inter-language link network, *WikiSYm 12: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, ACM, 2012 [Online]. Available: `http://doi.acm.org/10.1145/2462932.2462959`.

Choosing publics as a scholarly act: challenges for creators of digital editions at the point of inception

Elli Bleeker
University of Antwerp, DiXiT
elli.bleeker@uantwerpen.be

Aodhán Kelly
University of Antwerp, DiXiT
aodhan.kelly@uantwerpen.be

If creating a digital edition that meets the needs of scholars was not challenging enough in itself, editors need to make it both functional and attractive for their audiences as well. What does this imply for the traditional tasks of the editor? Elena Pierazzo argues that scholarly editors are typically the main public for their own editions. Her argument is confirmed by many current digital scholarly editions, which generally prioritise content above presentation and offer many complex tools for textual research that serve only the most advanced user. What other audiences should creators of editions attempt to engage with and how can they achieve a broader impact? The potential pedagogical value of using a digital edition in a classroom is one area that is regularly emphasised in such discussions. An editor needs to have a clear idea of the intended audiences right at the outset in order to structure their edition effectively. In other words, a predetermined public influences editorial decision-making about the shape and structure of an edition.

How do you create a digital edition that is potentially of interest to a larger community of users, while at the same time producing a product that is deemed to be 'scholarly'? Is there a conflict of interest between audiences that can be reached? Is it possible or even worthwhile to reach a multitude of publics, or should there be a solitary user group who are the intended focus? To what extent can the editors be in conversation with the potential audience to help shape the final output? From the point of inception of an edition project, editors must consider their specific (textual) material and the best modes to disseminate the work to their intended publics.

In this paper, we will discuss some of the issues and challenges listed above using a case in practice. We are currently at the outset of creating a new digital edition of a collection of Dutch language stories, Sheherazade (1932), by the Flemish author Raymond Brulez. By using a comprehensive and diverse set of surviving materials that form the basis of *Sheherazade*, we intend to represent its textual development on a detailed level. At this early stage we are confronted with a number of interesting conceptual and ethical challenges regarding publics and purposes. In this paper, we talk about the issues we face, the approach and model we hope to apply, and some preliminary observations we can make. In conjunction with this paper, we will give a related poster presentation in which we identify a number of practical and technical issues regarding the structure of edition. The challenge is to find a representation or combination of representations by which the edition can be both useful for scholarly purposes while also potentially providing wider pedagogical and societal value.

Combining quantitative and qualitative methods in digital analyses of literary style

J. Berenike Herrmann
German Department
Göttingen University
bherrma1@gwdg.de

This talk reports on a methodological agenda for digital analyses of literary style. Combining quantitative measures from stylometry and qualitative sample studies, I aim to fill the gap between large-scale methods of "distant reading" (e.g., Moretti 2000) and the examination of more intuitively visible patterns of language usage. Putting at the center of my study the prose of Franz Kafka embedded in a corpus of Modern German-written Literature (ca. 1880 - 1930, ca. 5 million words) I propose the following:

1. **Use quantitative measures to test literary hypotheses**: Specifically, I examine the question whether the claimed 'solitariness' of Franz Kafka's writing style (e.g., Oschmann 2010) is observable also in quantitative, formal terms (such as distance measures, vocabulary richness, keyword analysis, but also word and sentence length, cf. Stamatatos 2009; see also Craig et al. 2014). For example, a first application of distance measures confirms the idea that Kafka's writing is not very similar to his contemporaries (Herrmann 2013a).

2. **Use quantitative results as heuristics for ensuing qualitative style analysis**: The findings obtained in step 1 are not only suited to probe well-delineated hypotheses, but often also generate new kinds of data, allowing insights about what specific textual features may be indicative of a particular style. For example, a keyword analysis of Kafka's prose, combined with a manual post-hoc linguistic analysis indicates that Kafka appears to use more modal particles than other authors across genres (Herrmann 2013b). What is more, close inspection of features suggested that these are used in Kafka's text in a counter-intuitive way (Herrmann under revision).

3. **Feed qualitative analysis back to run finer-grained quantitative studies**: Findings obtained in step 1 and 2 may be fed back into quantitative style analyses, using measures that directly test new hypotheses about particular linguistic features. For example, I am currently building a corpus of Modern German literature that will be annotated for part-of-speech, in order to assess lexico-semantic variation across authors, text passages (including fictional characters), and time.

Step 1 is about the empirical examination of widely held scholarly findings, on formal and systematic terms and on a sizeable data basis. Step 2 allows making sense of the quantitative macro results on a micro-level, looking at stylistic phenomena that are closer to readers' experiences of style, thereby providing means of validation for

the quantitative measures. Step 3 then returns to the power of quantitative measures and hypothesis testing, however, with adjusted variables and stylistic measures. My talk will run through these three steps, reporting on my ongoing project on the digital analysis of Kafka's style, arguing that a combination of quantitative and qualitative methods in digital literary scholarship can be highly productive and valid, as well as thought-provoking.

References

Craig, H., Eder, M., Jannidis, F., Kestemont, M., Rybicki, J., & Schöch, C. (2014). 'Validating computational stylistics in literary interpretation', Paper presented at the *Digital Humanities Conference 2014*, Lausanne (`http://dharchive.org/paper/DH2014/Paper-857.xml`).

Herrmann, J. B. (2013a). 'Kafka among the authors. Stylometric analyses', Paper presented at the Expert Workshop "Stylometry@Kraków", Krakow, Poland.

Herrmann, J. B. (2013b). 'Computing Kafka – How keyness and collocation analysis help explain paradoxical style', Paper presented at the *International Herrenhausen Conference "(Digital) Humanities Revisited – Challenges and Opportunities in the Digital Age"*, Hanover, Germany.

Herrmann, J. B. (under revision). 'Läuse im Pelz der Sprache? Zu Funktionen von Modalpartikeln in narrativen (De-)Motivierungsstrategien bei Franz Kafka'. In M. Horváth & K. Mellmann (Eds.), *Die biologisch-kognitiven Grundlagen narrativer Motivierung*. Berlin: De Gruyter.

Moretti, F. (2000). 'Conjectures on world literature', *New Left Review*, vol. 1, pp. 54-68.

Oschmann, D. (2010). 'Kafka als Erzähler', In M. Engel & B. Auerochs (Eds.), *Kafka-Handbuch*, pp. 438-449. Stuttgart.

Stamatatos, E. (2009). 'A survey of modern authorship attribution methods', *Journal of the American Society for Information Science and Technology*, vol. 60(3), pp. 538-556.

Digital humanities: New toys for the boys?

Sally Wyatt
eHumanities KNAW
sally.wyatt@ehumanities.knaw.nl

The early days of computing (1950s-1970s) were characterized by a high level of gender equality. In what was then a new area of research and employment, there was no established gendered division of labour. This kind of engineering did not require physical strength but many tasks did require dexterity and attention to detail. The stereotype of the badly dressed, antisocial and male nerd/geek/hacker emerged somewhat later. The domination of computer science by men took time to take hold (Abbate, 2012; Ensmenger, 2010), and the view of digital technologies as a source of women's liberation or further oppression varied widely over the past 50 years (Wyatt, 2008).

The humanities as a whole (though there are important differences between fields) has always been much more open to women's participation (even if a glass ceiling continues to operate), and to gender as an object of analysis. So what happens when computer scientists and humanities scholars work together in the field of digital humanities? Concerns have been raised that digital humanities is simply a way of capitalizing the humanities, and enrolling it in an increasingly neo-liberal university environment. Some have complained that DH is not always sensitive to the epistemic traditions of the humanities. Others have gone so far as to suggest that digital humanities is a retreat from the critical tradition of the humanities (e.g. Zaagsma, special issue of BMGN, 2013; Grusin, 2014, drawing on debates held at MLA conferences). In this paper, I will pick up a related theme, and suggest that DH is a way of 'masculinising' the humanities, and I will do this in two ways. First, I will examine the available data regarding the participation of women (focusing on Dutch universities and DH centres). Second, I will address the more difficult question regarding whether or not the involvement of women makes any difference to the topics addressed, the methods used and the kinds of knowledge that are produced (Haraway, 1985; Harding, 1986). This will be done by analyzing key DH texts, using both close and distant reading techniques. This paper is intended to be provocative, in order to stimulate discussion and debate about the future direction of DH in terms of both women's participation and the role of gender as object and as critical lens.

References

Abbate, J. (2012) *Recoding Gender. Women's Changing Participation in Computing*. Cambridge, MA: The MIT Press.

Ensmenger, N. (2010) *The Computer Boys Take Over. Computers, Programmers, and the Politics of Technical Expertise*. Cambridge, MA: The MIT Press.

Grusin, R. (2014) 'The Dark Side of Digital Humanities: Dispatches from Two Recent mla Conventions', *Differences, A Journal of Feminist Cultural Studies* 25(1): 79–92.

Haraway, D. (1985) 'Manifesto for Cyborgs', *The Socialist Review* 80: 65–107.

Harding, S. (1986) *The Science Question in Feminism*, Cornell University Press, Ithaca.

Wyatt, S. (2008) 'Feminism, Technology and the Information Society: Learning from the Past, Imagining the Future', *Information, Communication & Society* 11(1): 112–131.

Zaagsma, G. (ed) (2013) *Special issue on Digital History, BMGN, Low Countries Historical Review* 128(4).

Digitization and Exogenesis

Wout Dillen
University of Antwerp
wout.dillen@uantwerpen.be

Dirk van Hulle
University of Antwerp
dirk.vanhulle@uantwerpen.be

Tom de Keyser
University of Antwerp
tom.dekeyser@uantwerpen.be

Ronan Crowley
Universität Passau
crowle01@gw.uni-passau.de

Vincent Neyt
University of Antwerp
vincent.neyt@uantwerpen.be

Within the field of genetic criticism, Raymonde Debray Genette coined the terms 'endogenesis' and 'exogenesis' to denote respectively the writing of drafts and the interaction with external source texts during the writing process. The proposed panel focuses on the ways in which exogenesis and its relationship with endogenesis can be given shape in a digital infrastructure. The case studies are the works, reading notes and personal libraries of James Joyce and Samuel Beckett.

'Catalogue these books': Digital Editions and the Digital Library
LSDI scanning represents the dominant economy of book digitisation. The proposed presentation looks to the intersections of these proliferating Big Text Data with digital editions of authors' manuscripts to ask what are the points of contact and friction that obtain between digital libraries and heavily curated, deeply encoded editions of archival materials? The case study animating the paper focuses on the commonplace notebooks that James Joyce compiled in the preparation of *Ulysses* (1922). Leveraging the practice of sustained text reuse that they reveal, Ronan Crowley will explore the challenges and possibilities afforded by competing or complementary scales of digital materiality.

Joyce's Library and the Extended Mind
A writer's library gives away more information about its owner than we would expect. Not only does it provide a glimpse into the novelist's "entire intellectual life at once" (Oram 2014, 2), the extended mind hypothesis suggests that a library and the information it captures is also situated within the extent of cognition (Van Hulle 2014, 151). In this respect, it has become a part of the mind that we are able to study and (if needed) reconstruct. This presentation suggests that creating a web framework for Joyce's extant and virtual library opens possibilities for the study of the writing process as well as for our research of the human mind.

Samuel Beckett's Library and Digital Exogenesis
Because personal libraries are invaluable resources for research into a literary work's writing process, the Beckett Digital Manuscript Project has recently developed a new

module: the Beckett Digital Library allows the user of our digital genetic edition to examine both his extant library and the reconstructed virtual library, and to assess the interpretative implications of the reading traces. The proposed paper shows how this module interacts with the other modules in our edition and suggests a model for the integration of digitized personal libraries in genetic editions.

References

Oram, Richard W. (2014) 'Writers' Libraries: Historical Overview and Curatorial Considerations', in Richard W. Oram (ed.), *Collecting, Curating, and Researching Writers' Libraries: A Handbook*, Plymoth: Rowman & Littlefield, pp. 1–28.

Van Hulle, Dirk. (2014) *Modern Manuscripts: The Extended Mind and Creative Undoing from Darwin to Beckett and Beyond*, London: Bloomsbury.

Editions and editors of Greek papyrological texts

Mark Depauw
KU Leuven
mark.depauw@kuleuven.be

Yanne Broux
FWO-Flanders, KULeuven
yanne.broux@kuleuven.be

Small questions can have large consequences. At the beginning of 2015, someone addressed the PAPY- list, the mailing list for all things papyrology, asking whether it was possible to look for editions of papyri after 1980. In Trismegistos (www.trismegistos.org), the year of edition was added in a systematic way for the about 60,000 (Greek) papyrological publications, but we realized this information was not easily accessible for the users of the online search engine. Moreover, the names of the editors of individual texts were still missing for thousands of papyrological publications.

After an appeal to the papyrological community, this situation was quickly mended through a collaborative effort on a Google Drive document. But work was not finished after this. If a text is published in an article in a journal, it is included in the volumes of the *Sammelbuch Griechischer Urkunden aus Ägypten* (SB). It is assigned a volume and a serial number, e.g. 'SB 14 11639' and is commonly referred to in this way. Here too, the information regarding the editors of texts was mostly missing in TM. Since the original publications can only be found through the entries in SB, it would have been a lot of work to enter them systematically. A new appeal was launched to the community, but probably because of the considerable effort involved, there was less response. Luckily, the editors of SB sent us a reverse index of publications and SB-numbers, and now our database of editors is more or less complete for Greek papyrological texts.

The idea then grew to use the data to study papyrological editing practices in a quantitative way, in the wake of Peter van Minnen's more impressionistic articles dealing with this subject (1993, 1994, 2007 and 2009). We supplemented the data of text editions with the material collected by the *Bibliographie Papyrologique*, by restructuring their author information (some 65,000 entries). This paper will present the results of this study of the 'century of papyrology'. We look into the chronological evolutions of text editions (both original and re-editions) and the rise (and decline?) of the papyrological community. For the practice of co-editing and co-publishing we also set up a network of individuals who collaborated and took centrality measures, edge weight and paths into account to get a feel of the dynamics of the papyrological community (e.g. the figure below).

References

van Minnen, P. (1993) 'The Century of Papyrology (1892-1992)', *Bulletin of the American Society of Papyrologists*, vol. 30, pp. 5-18.

Ibid., (1994) 'The Origin and Future of Papyrology. From Mommsen and Wilamowitz to the Present, from Altertumswissenschaft to Cultural Studies', A. Bülow-Jacobsen (ed.), *Proceedings of the 20th International Congress of Papyrologists*. Copenhagen, 23-29 August 1992, Copenhagen: Museum Tusculanum Press.

Ibid., (2007), 'The Millennium of Papyrology (2001-)', B. Palme (ed.), *Akten des 23. internationalen Papyrologen-Kongresses. Wien, 22.-28. Juli 2001*, Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Ibid., (2009), 'The Future of Papyrology', R.S. Bagnall (ed.), *Oxford Handbook of Papyrology*, Oxford: Oxford University Press.

Figure 6: Giant component of the network of co-editors of Greek papyrological texts.

Enabling personal authoring and narrative making: Sustainable reuse of digital objects for interactive teaching and learning tools

Catherine Emma Jones
CVCE, Luxembourg
catherine.jones@cvce.lu

Lars Wieneke
CVCE, Luxembourg
lars.wieneke@cvce.lu

Alexandre Genon
CVCE, Luxembourg

Frederic Reis
CVCE, Luxembourg

Frederic Allemand
CVCE, Luxembourg

Digital humanities and the corresponding transformation of paper based artefacts into digital objects as well as digital borne objects has triggered a move away from traditional models of academic publishing. It has led to the emergence of new forms of publishing for scientific knowledge, developing practices that are more attuned with web 2.0 technologies and the digital age. There has been a shift away from the simple metaphor of the traditional academic peer-reviewed paper recreated in electronic form towards platforms that are both machine and human readable and offer a more interactive and enhanced user experience.

These new and evolving publishing forms enable interactive and immersive user experiences, and provide greater transparency of the scientific research process, methods and data. Known as enhanced publications (ePublications) or rich internet publications (RIP), they encapsulate research knowledge whilst simultaneously providing mechanisms for describing, sharing, discovering, reuse and repurposing of the scientific content (Bechhofer, Roure et al., 2010).

With these advantages in mind, the CVCE.eu has been continuously improving its multilingual ePublication model and infrastructure, for both research and teaching and learning activities in European Integration studies. Seven key features underpin the CVCe's ePublication infrastructure (which comprises ePublications themselves and their constituent objects):

1. *Aggregations* – ePublications aggregate content derived from many different publishable objects which themselves are aggregations of individual resource objects such texts or photographs.

2. *Identity* – all objects and ePublications have an individual and persistent identfier in the form of a unique URL.

3. *Traceability and transparency* – CVCE users must be able to trace the steps undertaken by the researcher through transparency in research methods to produce the ePublication.

4. *Metadata* – without descriptive machine-readable metadata the ePublications and their constituent objects could not be reused or repurposed effectively.
5. *Reuse* – Enabling reuse of the object in another context but with the same content for example as part of user derived ePublications.
6. *Repurpose* – Changing the way the object is used for example integrating it within a timeline of a map.
7. *Sustainability* – ensuring long-term access to objects (in relation to copyright restrictions) and through the use of persistent identifiers.

The result is a set of historical narratives describing European integration themes and topics as told by a researcher(s). The ePublications are aggregations of diverse research objects including: historical documents, press articles, photographs and other multimedia material, each with a set of descriptive metadata, publishable metadata and a unique, persistent identifier.

The ePublication platform is challenged by growing user demand for personalisation, collaboration, participation and sharing features commonplace within social technologies. To fulfil these requirements, the CVCE needs to develop new teaching and learning tools harnessing the power of repurposing and reuse of all unique objects, as well as digital tools that enable users to personalise content from across our virtual research infrastructure. The resulting Digital Toolbox facilitates customisation and personalisation of content with the mid-term goal to provide a personalised research infrastructure where users can create, author, collaborate, comment and share self-authored ePublications.

Figure 7: Screenshot of the MyPUblication authoring tool interface.

This paper describes the beta version of the Digital Toolbox which concentrates efforts on the design and implementation of the MyPublications authoring tool (see figure). The MyPublications tool enables users to write ePublications on topics of

interest based on personalised research, teaching and learning goals by organising and restructuring CVCE resources (documents, images, treaties, photographs etc). The tool includes a presentation viewer, a simple interface where users can sequentially step through an ePublication (like a book!) and incorporates a share link feature. With such new interactive tools, the hope is to contribute to a richer educational experience through active and constructive learning processes.

References

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. *The Future of the Web for Collaborative Science (FWCS 2010)*, April 2010, Raleigh, NC, USA (`http://precedings.nature.com/documents/4626/version/1`).

Europe's Beginnings through the Looking Glass: Publishing Historical Documents on the Web Using EVT

Roberto Rosselli Del Turco
Università di Torino
roberto.rosselli delturco@unito.it

Florentina Armaselu
CVCE
florentina.armaselu@cvce.eu

Lars Wieneke
CVCE
lars.wieneke@cvce.eu

Chiara Di Pietro
Università di Pisa
dipi.chiara@gmail.com

Raffaele Masotti
Università di Pisa
raffaele.masotti@gmail.com

EVT (Edition Visualization Technology) [1] is a software aimed at the creation of image-based web editions of TEI P5 encoded texts. It is a light-weight, open source tool specifically designed to simplify the production of digital editions, freeing the scholar from the burden of web programming and enabling the user to browse, explore and study digital editions by means of a user-friendly interface, providing a set of tools (zoom, magnifier and hot-spots for manuscript images, an internal search engine for the edition texts) for research purposes.

Everything is created around the data and the encoded text itself: by applying a single style sheet to the TEI XML file that contains the whole transcription of a document, an XSLT 2.0 transformation chain is started that results in a web based application – a mix of HTML 5, CSS3 and JavaScript – that can be easily shared on the Web. Besides presenting the digital scans of the original manuscript (if available) linked to the corresponding text of the edition, the software provides a bookreader visualization mode if double side images are supplied.

EVT was born in the context of a specific use case (the Digital Vercelli Book project, whose first version has been available online for about a year [2]), but it is now being used to publish another digital edition, that of the Codice Pelavicino manuscript [3]. The need to adapt it to different types of documents has led to a revision and expansion of the underlying code to make it more flexible and suitable for many different types of TEI-encoded texts. With this proposal we want to demonstrate the flexibility and re-use of EVT by applying it to an edition of diplomatic documents to be published on the CVCE's Web site [4]. We will discuss both the technical implications of its application as well as methodological consequences.

The intended digital edition we plan to build using EVT is based on XML-TEI P5 bilingual (French, English) documents of the W.E.U. (Western European Union), concerning armament production, standardisation and control in the period 1954- 1982. The corpus was selected from the Luxembourg National Archives, W.E.U. collection, and implied: OCR processing with ABBYY FineReader (one image file per typewritten page), Microsoft Word styling and OxGarage [5] conversion to XML-TEI P5, as well

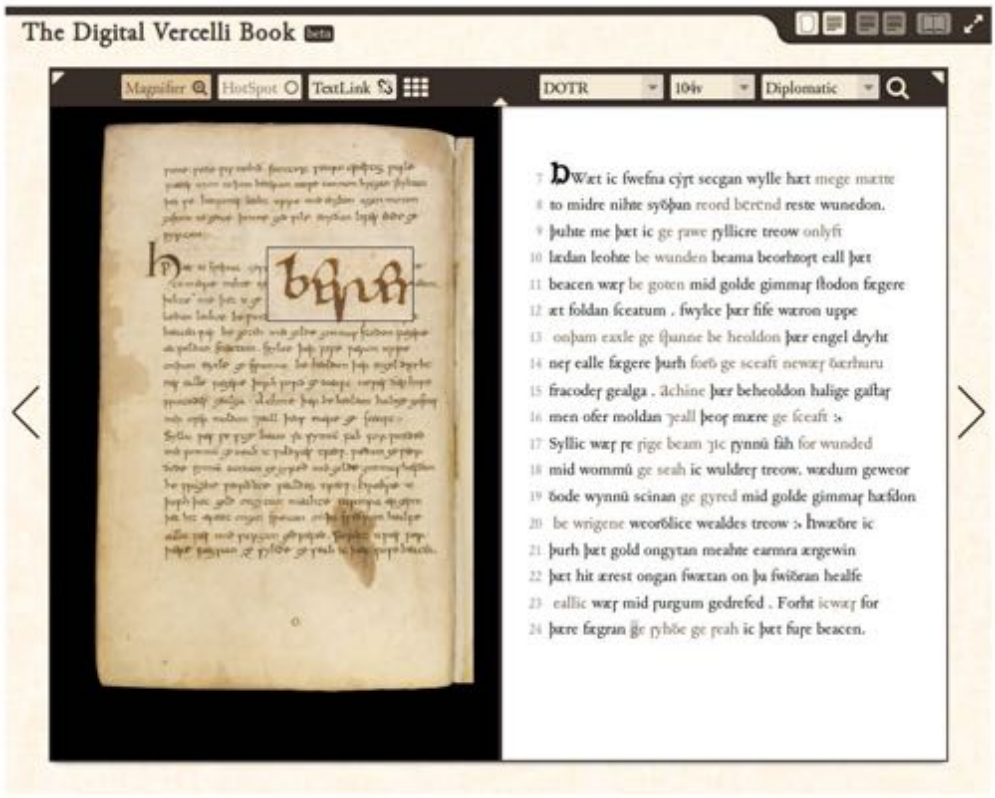


Figure 8: The Digital Vercelli Book edition using EVT.

as semi-automatic enrichment by XSLT 2.0, Named Entity Recognition with GATE [6] and manual annotation.

Several types of documents compose the corpus: meeting minutes, notes from the Secretary-General, memoranda and studies. The encoding includes metadata (title, author, document reference, copy number, version, date of availability, confidentiality status); structural markup for headers/footers, sections/subsections, paragraphs, line breaks; content-related annotations (discourse of institutional/country representatives, names of persons, organisations, etc.).

A digital edition will imply features for browsing and searching the collection, as well as side-by-side visualisation of the transcribed pages and facsimile images. The presentation would address technical issues (adjustment of the EVT framework to generate the Web edition from the XML-TEI corpus) and more general questions (to what extent the XML-TEI encoding and associated technologies may support enhanced 'mirror-like' digital representations of the original documents). The presentation will end with a short demo of the digital edition, based on the documents that have been digitized and encoded.

Notes

[1] EVT home page: `https://sourceforge.net/projects/evt-project/`.

[2] Digital Vercelli Book beta edition: `http://vbd.humnet.unipi.it/beta/`.

Figure 9: The digital edition of the Codice Pelavicino manuscript.

[3] Codice Pelavicino digital edition: `http://labcd.humnet.unipi.it/`.
[4] CVCE (Centre Virtuel de la Connaissance sur l'Europe): `http://www.cvce.eu/`.
[5] OxGarage: `http://www.tei-c.org/oxgarage`.
[6] GATE (General Architecture for Text Engineering): `https://gate.ac.uk/`.

Event-based object identification

Alina Saenko
PACKED vzw
Expertisecentrum
Digitaal Erfgoed
alina@packed.be

In March 2015 PACKED vzw started the 'Event-based object identification' project, exploring how information in collection management systems about the lifecycle of artworks can be transformed into rich machine-readable data suitable for digital humanities research. This project is part of a series projects funded by the Departement Culture, Youth, Sports and Media of the Flemish Government on the practical implementation of concepts such as persistent identification, Open Data and Semantic Web in the cultural heritage sector.

Museums record information about the lifecycle of an artwork, which is relevant for collection management purposes, i.c. acquisitions, loans, restorations etc. Information about production, ownership, restoration and exposition is easily accessible for museum staff through the collection management system, but hardly for any external users. This information however provides a rich context around an artwork that may provoke new ideas and possibilities for digital humanities research. Museums should ask themselves how they could unlock this potential.

Initiatives such as CIDOC-CRM and LIDO introduced the concept of the 'event' in museum documentation practice as an instrument to identify and cluster factual information about the life of an artwork. Events are generic containers that reveal what happened at a particular time and place to a work of art. In practice, information on an event in a life of an artwork lies deeply embedded in collection management oriented data structures, often spread over different software systems. The current project will look into the possibilities for extracting this information from these data structures and systems and find a sustainable way to make it available for research.

Creating large sets of events, related to one or a group of artworks, may introduce new research perspectives that benefit from the rich spatial and temporal dimensions of events as well as the networks of people that are linked to these events. The project will work with rich contextual data of 35.000 artworks throughout the collections of seven Flemish art museums and aims at finding concrete solutions how these museums can use this data to nurture the research community.

During DHBenelux 2016 the first results of this project will be presented:

1. What is the influence of using persistent URI's to identify data about artworks? (presentation of the results of the previous 'Persistent identification' (2013- 2104) project as a starting point of this project)

2. Where do museums store data about events and how can it be extracted and reused?

3. How should events and data about events be identified online? Which external authorities should be used to identify when–where–who data about events?
4. What visualisations and research opportunities will be possible after the normalisation of data about events?

Examining the Bomb Sight augmented reality app for exploring location based historical data

Catherine Emma Jones
CVCE, Luxembourg
catherine.jones@cvce.lu

Dan Karran
GeoBits Ltd

Patrick Weber
Location Insights Ltd

The digital age has impacted how historical artefacts such as texts, maps, personal narratives and pictures are digitally recorded, enhanced, enriched, analysed, utilised and disseminated. Digitised historical artefacts and associated digital tools and methods provide new opportunities to support researchers, teaching and learning communities by offering contextually situated information outside of libraries and archives. The result is more widely accessible archival data, combined with other forms of historical information, conflated within novel, interactive and user-friendly interfaces across a range of devices. With the now almost ubiquitous nature and use of geo-enabled technology and data, driven by the likes of Google Maps, Open Street Map etc. we are now in the era of digital interactive maps, used across a range of application areas including digital humanities projects. In this paper we present a case study describing the design and development of an augmented reality application for users of Android mobile devices. The application, Bomb Sight, enables users to view historical bomb census maps from WW2, for London during October 1940 to June 1941, previously only accessible in the map reading room of The National Archives, Kew, London, together with an overlay of the bomb locations on top of the present day London cityscape.

Harnessing the power of the modern (Android) smartphone platform, the Bombsight AR app makes use of geo-location technology and the smartphone camera to overlay digital objects, the historical bomb census information, over the present day real world, see figure 1. The augmented reality view shows the user markers hovering over where bombs fell, scaled to show closer locations with larger markers and smaller ones for those further away. For added context, street name labels issued from the historical records are added to each marker where available. By clicking on the marker, the user can view further information about the bomb type, and how far away it is from the users location. There is an associated radar plot highlights the bomb location in in relation to others.

The project is a successful case study how to open up and make use of digitised archive data in the context of 20th century history, engaging new audiences through augmented and contextualised historic records. The AR application successfully projected historical information into the present world, situating historical information seamlessly in relation to time and place, offering a window into the past through a smartphone display.

Figure 10: Augmented reality view of the `www.bombsight.org` app.

Acknowledgements Bomb Sight was created by a collaboration between Dr Catherine Jones, (formerly University of Portsmouth, currently CVCE.eu, Luxembourg) and The National Archives (Andrew Janes) with development of the web application by Dr Patrick Weber of Location Insights Ltd and the mobile app by Dan Karran of Geobits Ltd. Jasia Warren created the design of both tools.Its creation was funded by JISC-Joint Information Service Committee under Strand C: Clustering Digital Content of their Content Programme 2011-13.

From lighthouse to the moon: a guiding light to the corpus of Jules Verne

Nicholas J. Hayward
Loyola University Chicago
nhayward@luc.edu

George K. Thiruvathukal
Loyola University Chicago
gkt@cs.luc.edu

Introduction

The Verne Digital Corpus (VDC) [1], which focuses upon the work of Jules Verne, will include original French language editions and their myriad, often questionable, English language translations.

Publication of Digital Scholarly Editions

As part of the development of Woolf Online (WO) [2], we developed an extensible development and publication framework, called 'Mojulem' [3], for editing, publication, and visualisation of digital scholarly editions. This development continues with the VDC4, which focuses upon the work of Jules Verne, including French language editions and their myriad, often questionable, English language translations. Mojulem allows us to build on the concept of 'knowledge sites', as suggested by Peter Shillingsburg (Shillingsburg 2006), supplementing a core publication framework with modules such as OCR, editors, and image viewers. Development followed the need for four initial core structures. These structures include CorPix, CorTex, CorCode, and CorForm, which have now been supplemented with CorAssess for the VDC.

Why Verne?

After WO, we chose to focus upon the corpus of Jules Verne, including French language editions and English language translations. We have begun collecting, collating, and preparing digitised copies of as many digitised editions as extant online. We have also begun digitising early editions to provide an ever-growing dataset of Verne material. The nature of early English language translations of Verne's editions is a continuing source of frustration for those interested in his works. His early categorisation as a predominantly children's author in English language countries, unlike the publishing by Hetzel, coupled with early restricted access to original French language editions, simply compounded the issue.

The corpus of Jules Verne offers an interesting opportunity for literary and contextual analysis coupled with data processing and automated analysis. The myriad existing digitised English language translations, including US and British variant editions, often more prevalent than their counterpart French editions in our current digitised corpus, allows us to examine the development of those texts by comparing agreements, disagreements, omissions, and continuing revisions in said translations since a novel's first edition. We are hoping to use this analysis to filter the noise of years of collective translations to collate a unified English translation for each French language edition. We will then be offering a comparison of French language edition against a filtered, collated English language edition. This will allow further consideration of the

requisite merits of the English language translation directly juxtaposed to the original French language edition.

Conclusion

With the addition of CorAssess, we are now beginning to address additional issues with the publication, transmission, and development of texts. We are also testing, and proving, the viability of Mojulem beyond the WO project. The corpus of Jules Verne provides a particularly fascinating opportunity to test these cores, and provide a resultant conflated, English language translation per extant French language edition. This paper will briefly introduce the Mojulem framework and its initial four cores, and detail the ongoing developments to augment this work with the above new work on the corpus of Jules Verne, and the ongoing VDC.

Notes

[1] Initial development project site is available at the following URL: `http://www.scifidocs.com/`.
[2] Project site is available at the following URL: `http://www.woolfonline.com`.
[3] Our current test framework for the Woolf Online project, as of 6th April 2015, can be viewed at the following URL: `http://dhdev.ctsdh.luc.edu/projects/edfu/`.
[4] Initial development project site is available at the following URL: `http://www.scifidocs.com/`.

References

Goodman, N. 1976, *Languages of Art*, 2nd edn. Hackett Publishing Company, Boston.
Hayward, NJ, Shillingsburg, PL & Thiruvathukal, GK 2013, *Woolf Online*. Available from: `http://www.woolfonline.com` [6 April 2015].
Hayward, NJ 2015, *SciFi Docs*. Available from: *http://www.scifidocs.com* [6 April 2015].
Shillingsburg, PL 2006, *From Gutenberg to Google: Electronic Representation of Literary Texts*,1st edn. Cambridge University Press, Cambridge.

GIStorical Antwerp: a micro-level data tool for the study of past urban societies, test-case: Antwerp

Iason Jongepier
University of Antwerp
iason.jongepier@uantwerpen.be

Ellen Janssens
University of Antwerp
ellen.janssens@uantwerpen.be

Space never lies. Urban societies, historians agree, are distinct through their size, density and the spatial relationships that shape social interactions, which in turn are at the heart of urbanity. Yet, hitherto few studies have managed to integrate spatial data into a comprehensive database system that allows to map the complex web of social relations. Moreover, most GIS-based studies tackle cities at an aggregate level, while most social decisions and data are created at a micro-level. The major goal of the Gistorical Antwerp project is to build a micro-level data tool for historical Antwerp, but transferable to other cases as well, constructing series of vector maps, linked to a variety of historical geo-data. By creating an historical GIS for urban history, we argue, amazing synergies can be reached through the integration of various sets of data into one spatial database: future archaeological research will no longer be excavating the remains of an anonymous stone house, but, at the first onset of the enquiry, it will know that in the nineteenth century this site was the workplace of, for example, a tanner. And a historical study on, for example, parish-charities can be completed with various formerly inaccessible datasets on housing, occupational activities, membership of guilds, activities on the real estate market, marriage alliances etc., using spatial location to integrate the most diverse datasets.

The development of GISTorical Antwerp necessarily starts in the (long) nineteenth century, when the first modern cadaster was implemented. Based on this cadaster a spatial base layer of over 9,000 plots was already digitalized. These plots will be linked to two large and very pivotal datasets for the history of nineteenth century Antwerp, the commercial almanachs (directories of industrial and commercial activities) and the building/environmental permits. During the nineteenth century, the urban landscape and society were complete reshuffled by both active town planning and the far-reaching impact of industrialization; mass migration; transport (r)evolutions; social polarization; the breaking-up of old solidarities (e.g. the urban guild system), and their replacement by new ones; and specifically for Antwerp, the rise of the city as major European harbor. The integrating of both data-sets in GIStorical Antwerp, not only allows to analyze changing topographies of commerce and industry, but also to measure how these topographies incorporated and steered the above mentioned transformation processes. Once constructed, GIStorical Antwerp immediately allows to address major research questions on the social, environmental and economic history of the nineteenth century town, for the first time at the level of the individual household. Thanks to this unique combination of a critical mass of already existing databases and a clear focus on households as major agents of social behavior, we believe, this cutting-edge tool will result in a truly innovative “integrated social history” of the city.

Ghost in the Machin...ima: Reflecting on Machinima's Ability of Going Mainstream

Aris Emmanouloudis
University of Amsterdam
ademmano25@yahoo.gr

Fanfiction is a term closely associated with today's emergent media culture. One of the most popular ways of fans creating their own content is engagement with the practice known as "Machinima". Machinima means the creation of videos by use of computer graphics borrowed from other fields, most often video games. Kate Fosk (2011) has claimed that the area of machinima use is evolving and abandoning its juvenile roots, becoming more mature thus enabling it to stand next to traditional media practices, such as filmmaking. On the other hand, Friedrich Kirschner (2011) argues that machinima is incapable of such extents, being limited by its production tools. This presentation, after tracing a brief history of the machinima practice, attempts to answer if machinima is just something appreciated among certain fan circles bound to the game culture, or something bigger appealing to a broader audience and capable of achieving more.

The method that I will follow is a comparison between the main arguments that support and those that challenge machinima's ability of going mainstream, and drawing by relevant literature and examples, I will attempt to reach to a conclusion on whether machinima is an art form ready for the mainstream world or not.

References

Fosk, K. (2011) 'Machinima Is Growing Up', *Journal of Visual Culture*, vol. 10, April, pp. 25-30.
Kirschner, F. (2011) 'Machinima's Promise', *Journal of Visual Culture*, vol. 10, April, pp. 19-24.

Globalising a Pan-European Common Names Webservice @ the Cloud: 5 steps towards Linked Open Data Humanities

Eveline Wandl-Vogt
Austrian Academy of Sciences
eveline.wandl-vogt@oeaw.ac.at

Heimo Rainer
Naturhistorisches Museum Wien
heimo.rainer@nhm-wien.ac.at

Thierry Declerck
Deutsches Forschungszentrum für Künstliche Intelligenz
declerck@dfki.de

The digital media is changing the way we think, and – even more quickly – the way we work and with that access and navigate in digital surroundings and make use of data and tools [1,2]. Language is the tool mostly used to access the digital world. What is common for standard language gets challenging when working with non-standard language such as dialectal or historical data. In this paper the authors discuss a Globalisation of an existing knowledge resource, namely the Common Names Webservice, developed by the Naturhistorisches Museum Wien (nhm) in the framework of the European Project of OpenUp! [13] in conjunction with the Austrian Academy of Sciences, Austrian Centre for Digital Humanities on the example of the Database of Bavarian dialects in Austria (DBÖ). The authors demonstrate a step-to-step integration of new resources in their model on the example of the Dutch Pland Database.

The authors focus on 5 ways to improve this common service and – in doing so – several humanities fields and discusses how this improvements vice versa have a positive input on digital humanities in general:

1. Convert a dictionary data | common names collections into machine readable format and publish it in the LOD [3]. The authors exemplify an existing model for lexicographic data in the cloud, namely the model of the ontology-lexica working group [6]. Furthermore, the model used for common names service is discussed and described.

2. Create an interactive service rather than an object [4]. The authors show, how they build an interactive webservice, dealing with linguistic and botanical data. They discuss the added value for the connected team-members and embed the results into their specific workflows. They summarize with suggestions for added value for interdisciplinary research infrastructures.

3. Enrich the service with collaborative expertise.The authors discuss on the example of the Collaboration of the Botanists at the Naturhistorisches Museum Wien (embedded into the European Project "OpenUp! Opening Up Europes Natural History Museums for Europeana") and the Lexicographers (embedded into the European Network for electronic Lexicography (ENeL) [11] and the European Research Infrastructure DARIAH.EU) [12] at the Austrian Academy of Sciences. Furthermore, they point out the added value established by the collaboration of

botanists and lexicographers making use of language technology and semantic web technologies as well as linked (open) data technologies.

4. Create several paths for access and analytics [5]; keep open minded for human beings, so keep all senses and disciplines in mind to assure best ways of data reusability. The authors give several examples for re-using of those data:
   (a) Europeana [7]
   (b) Wikipedia [8]
   (c) Babelnet [9]
   (d) Agrovoc [10]
5. Connect and globalize, data, services as well as people. On the example of a new project of a European Plant Names Thesaurus developed within the framework of the COST ENeL action, the authors discuss the possibilities of interdisciplinary collaboration. On the example of the Common Names Service, the author discusses the vision of worldwide virtual research environments, focussed on certain issues, yet embedded into global infrastructures. They offer opportunities to join in and be part of.

Concluding, added values of Linked Data Humanities on the example of interdisciplinary controlled Vocabularies will be presented for discussion and open feedback.

References

[1] Carr, Nicolas "Surfen im Seichten. Was das Internet mit unserem Hirn anstellt." München 2013.
[2] Nentwich, Michael, König, René "Cyberscience 2.0. Research in the Age of Digital Social Networks." Wien 2012.
[3] Wandl-Vogt, Eveline, Declerck, Thierry "Mapping a dialectal dictionary with Linked Open Data". In Kosem, Iztok (et al., Eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*, 17-19 October 2013, Tallinn, Estonia: 460-471 (http://eki.ee/elex2013/proceedings/eLex2013_32_Wandl-Vogt+Declerck.pdf [accessed: 15.10.2014]).
[4] Tasovac, Toma "Reimagining the dictionary, or why lexicography needs Digital Humanities." *Digital Humanities* (2010): 1-4. (http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-883.pdf [accessed: 15.10.2014]).
[5] Theron, Roberto, and Laura Fontanillo "Diachronic information visualization in historical dictionaries." *Information Visualisation* (2013).
[6] Ontology-Lexica Working Group: http://www.w3.org/community/ontolex/ [accessed: 15.10.2014].
[7] Europeana. Think Culture (http://www.europeana.eu/ [accessed: 15.10.2014])
[8] Wikipedia: http://de.wikipedia.org/wiki/Wikipedia%3AHauptseite[accessed:15.10.2014]
[9] Babelnet: http://babelnet.org/ [accessed: 15.10.2014].
[10] Agrovoc: http://aims.fao.org/agrovoc [accessed: 15.10.2014].
[11] COST ENeL: COST IS 1305: European network of electronic Lexicography: www.elexicography.eu [accessed: 15.10.2014].
[12] DARIAH.EU: European Research Infrastructure for the Arts and Humanities: www.dariah.eu [accessed: 15.10.2014].
[13] OpenUp! Opening Up the Natural History Heritage for Europeana: http://open-up.eu/ [accessed: 15.10.2014].

God does not sing. Identification of participants in Psalm 75

Christiaan Erwich
VU University Amsterdam
christiaan.erwich@gmail.com

Wido van Peursen
VU University Amsterdam
w.t.van.peursen@vu.nl

What voices can be heard in the Psalms? A great challenge of reading the poetry of the Psalms is the identification of participants. The major cause of this problem is a continual shift in person, number and gender (so-called PNG-shifts) in the text. To account for these shifts, scholars have always used an intuitive literary analysis. However, this exegetical tradition has led to diffuse and ad hoc explanations of: 1. Who the speakers are in the Psalms: a psalmist (Uchelen 1977), priest (Kraus 1978), king (Tate 1990), prophet (Ridderbos 1958; Terrien 2003), or God?; and 2. What the setting and genre is of these texts: e.g., liturgy or oracle? The current scholarly approaches therefore lack a systematic registration of patterns of PNG-shifts, as well as a methodologically adequate analysis.

To make a start with a more systematic analysis of PNG-shifts, we analysed Psalm 75 as an interesting case study. In this pilot project we used the annotated database of the Hebrew Bible prepared by the *Eep Talstra Centre for Bible and Computer* to help formulate an answer to one of the central interpretation problems of Ps. 75: the separation of human and divine speech. A linguistic four-step analysis was made. Firstly, data on person, number and gender of all verbs, pronomina and suffixes were gathered from the ETCBC-database and categorised. Secondly, since, the ETCBC-database does not contain identifications of participants, participants were be identified manually by relating them to the most important persona in Ps. 75. Thirdly, we searched for patterns such as the 1pl.-1sg. shift as in "We [1pl., the community] give thanks to thee [God]... I [1sg, God] will judge with equity". Fourthly, in order to formulate new ideas about the coherence and categorisation of the text of Ps. 75, we demarcated direct speech sections: e.g. an oracle by God is often not introduced by such but should be inferred by a change of speaker [God] into addressee [human].

The experiments did not in all respects meet the expectation and we learned valuable lessons that we will use in future research. Since the past project was only a pilot, we plan to do a more exhaustive analysis in the near future, taking into account the lessons we have learnt.

References

Hossfeld, F-L. (2000) *Psalmen 51-100*. Herders theologischer Kommentar zum Alten Testament. Freiburg: Herder.

Kraus, H-J. (1978) *Psalmen II 60-150*. Biblischer Kommentar Altes Testament Bd. 15. Neukirchen-Vluyn: Neukirchener Verlag,.

Ridderbos, J. (1958) *De Psalmen II*. Commentaar op het Oude Testament. Kampen: Kok.

Tate, M. E. (1990) *Psalms 51-100*. Word biblical commentary vol. 20. Waco: Word Books.

Terrien, S. L. (2003) *The Psalms: strophic structure and theological commentary*. Grand Rapids: Eerdmans.

Uchelen, N. A. van (1977) *Psalmen. Dl. II, (41-80)*. De prediking van het Oude Testament. Nijkerk: Callenbach.

HEEM, a Complex Model for Mining Emotions in Historical Text

Inger Leemans
VU University Amsterdam
i.b.leemans@vu.nl

Janneke M. van der Zwaan
Netherlands eScience Center, Amsterdam
j.vanderzwaan@esciencecenter.nl

Isa Maks
VU University Amsterdam
e.maks@vu.nl

Erika Kuijpers
VU University Amsterdam
erika.kuijpers@vu.nl

Recently, emotions and their history have become a focus point for research in different academic fields (Matt and Stearns 2013; Plamper 2015; Boddice 2014). At the same time, sentiment analysis and opinion mining have become important research areas, both within and outside academia. So far, however, many techniques are aimed at fitting relatively simple emotion models (positive/negative emotion, or limited sets of at most of six or seven 'basic' emotions (e.g., 'anger', 'disgust/contempt', 'fear', 'interest', 'joy', 'love', 'sadness', and 'surprise') (Buitinck et.al. 2015). In addition, these simple models are almost exclusively applied to contemporary, and/or web-based texts (e.g., Strapparava and Mihalcea 2008; Yang, Lin, and Chen 2007).

However, simple emotion models are not sufficient for Digital Humanities scholars who are interested in research questions about changes in emotional expressions over time. Answering these questions requires more complex, historically motivated emotion models. Also, because historical (literary) text differs significantly from contemporary, and/or web-based text, e.g., with respect to spelling, it is not clear to what extent modern approaches to emotion mining work on historical text. These two issues are addressed in this paper.

This paper presents a new model for emotion mining, resulting from the research project 'Embodied Emotions'. This project aims: 1. to trace historical changes in emotion expression and in the embodiment of emotions, and 2. to develop methods to trace these changes in sizeable corpuses of digitized texts. To meet these challenges, we have developed the Historic Embodied Emotion Model (HEEM), built on a test case of 29 plays Dutch language theatre plays written between 1600 and 1800 and annotated manually with HEEM labels for emotions and body terms. In this paper we present this model and compare it with other sentiment mining techniques, e.g. off the shelf linguistic analysis software LIWC (Linguistic Inquiry and Word Count) (Pennebaker 2001) and a version of LIWC that has been adapted for the analysis of Dutch historical texts

We will conclude that, although different forms of sentiment mining have their value and use, for emotion mining and analysis of embodied emotions in historical texts, HEEM sets a new standard.

References

Boddice. R. (2014) 'The affective turn: historicising the emotions'. In: C. Tileag and J. Byford (ed.), *Psychology and history: Interdisciplinary explorations*, pp. 147))156. Cambridge: Cambridge University Press.

Buitinck, L., Van Amerongen, J., Tan, E. and De Rijke (2015), M. 'Multi-Emotion Detection in User-Generated Reviews', *ECIR 2015: 37th European Conference on Information Retrieval*.
Matt, Susan J., and Peter N. Stearns (eds.) (2013) *Doing Emotions History*. University of Illinois Press.
Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001) *Linguistic Inquiry and Word Count: LIWC2001*. Mahwah, NJ: Erlbaum Publishers (`www.erlbaum.com`).
Plamper, Jan. 2015. *The History of Emotions: An Introduction*. Oxford: Oxford University Press.
Strapparava, C. and Mihalcea, R. (2008) 'Learning to identify emotions in text', *Proceedings of the 2008 ACM symposium on Applied computing*, pp.1556–1560.
Yang, C., Lin, K.H. and Chen, H.H. (2007) 'Emotion classification using web blog corpora', *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 275–278.

Historical data exploration: Amsterdam's creative landscape, 1600-present

Leonor Álvarez Francés
Leiden University, University of Amsterdam
l.alvarez.frances@hum.leidenuniv.nl

Rosa Merino Claros
University of Amsterdam
r.merinoclaros@uva.nl

Harm Nijboer
University of Amsterdam
h.t.nijboer@uva.nl

Julia Noordegraaf
University of Amsterdam
j.j.noordegraaf@uva.nl

Claartje Rasterhoff
University of Amsterdam
c.rasterhoff@uva.nl

In this panel session members from the research program Creative Amsterdam: An E-Humanities Perspective (CREATE) present and discuss their research. CREATE researchers collect and enrich digital data on the various cultural sectors of Amsterdam, link existing datasets, and develop novel search and analysis tools (http://achi.uva.nl/create). By doing so, they promote an infrastructure for combining, analyzing, and visualizing existing datasets in a network that exposes their relations and interdependencies. In this session, the central aim is to examine how the interaction between historians and their data can be expanded and enhanced.

During the last two centuries, historians have built a treasure trove of information on cultural consumption, production and distribution. But how to turn this into data? And how to access and analyze it in ways that can reveal new relationships, patterns and research questions on the development of Amsterdam as a creative city? In this session we address these questions by means of three cases: theatre (17-18th centuries), heritage (17th century) and cinema (20th century). Each case highlights a different aspect of the promises and challenges of data exploration in historical research, but they all center on the question of how datasets and analytical tools can meet the needs of historians. After a brief introduction on this central theme, the three cases will be presented (5-10 minutes) and central discussion questions will be posed to the audience.

- The ONSTAGE project exemplifies processes of digitizing historical data. ONSTAGE will contain data on the repertoire, the revenues, and the networks of the Theatre of Amsterdam between 1638 and 1772. Until recently, this information was dispersed, and available only in the form of partial studies and printed datasets. In ONSTAGE project shows how integrating and digitizing historical data fosters new research questions.

- The CANAAN project showcases the meeting of two datasets, originally designed for different purposes. The collection database of the Amsterdam Museum contains metadata on more than 90,000 objects. The ECARTICO database

(UvA) contains biographical data on Dutch artists from the early modern period. Aligning the two datasets allows for advanced querying across the two resources.

- The CINEMA project illustrates how existing datasets may acquire new functionality. The Cinema Context database contains information on theaters, people, companies, and visitors of Dutch cinemas from 1896 to the present, and is explicitly structured for research purposes. This project explores how additional functions can be developed in order to allow for, for instance, historical network analysis.

How Can Language Technology Fight Against Language Death?

Ivett Benyeda
Hungarian Academy of Sciences
benyeda.ivett@nytud.mta.hu

Eszter Simon
Hungarian Academy of Sciences
simon.eszter@nytud.mta.hu

Péter Koczka
Hungarian Academy of Sciences
koczka.peter@nytud.mta.hu

According to the UNESCO Atlas of the World's Languages in Danger (Moseley 2010) there are 646 definitely endangered, 528 severely endangered and 576 critically endangered languages. The European Union places great emphasis on the preservation of linguistic diversity, which means that it is a common aim to support endangered languages and prevent language death.

The digital revolution of our era has a dramatic impact on nearly all aspects of society. Language communities are most sensitive to and therefore most affected by new paradigms in communication technology (Simon et al. 2012). According to Kornai (2013), a language is digitally viable only to the extent it produces new, publicly available digital material. Language death implies loss of function, entailing the loss of prestige, and ultimately the loss of competence. In this context, language technology aspires to become an enabler technology that helps people to collaborate, conduct business and share knowledge regardless of language barriers (Simon et al. 2012). However, cutting-edge technologies are typically available only for widely-spoken ('thriving') languages (Kornai 2013).

In this presentation, our aim is to present an ongoing project whose objective is to produce digital material for the following endangered Finno-Ugric (FU) languages: Komi-Zyrian, Komi-Permyak, Udmurt, Meadow and Hill Mari and Northern Sami, helping them in the process of revitalisation. To achieve our goals, we collect parallel, comparable and monolingual texts for the mentioned small FU languages and for thriving languages that are of interest to the FU community: English, Russian, Finnish and Hungarian. We generate proto-dictionaries for the FU–thriving language pairs and will deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary. See the workflow in the figure below: the green-coloured steps have already been conducted, the orange ones are under development, while the red one indicates the final step which will be taken in the last phase of the project.

First, we created proto-dictionaries from already existing, digitally available dictionaries, from Wiktionary by using parsing and triangulating methods (Ács 2013), and by extracting title pairs from Wikipedia. For extracting translation candidates from parallel corpora, we use the HunDict tool [1]. We are experimenting with methods to extract real parallel sentences from comparable data, which can then be used as input material for generating new word pairs. Several similarity measures can also be used for extracting translation candidates directly from comparable corpora. The difficulty

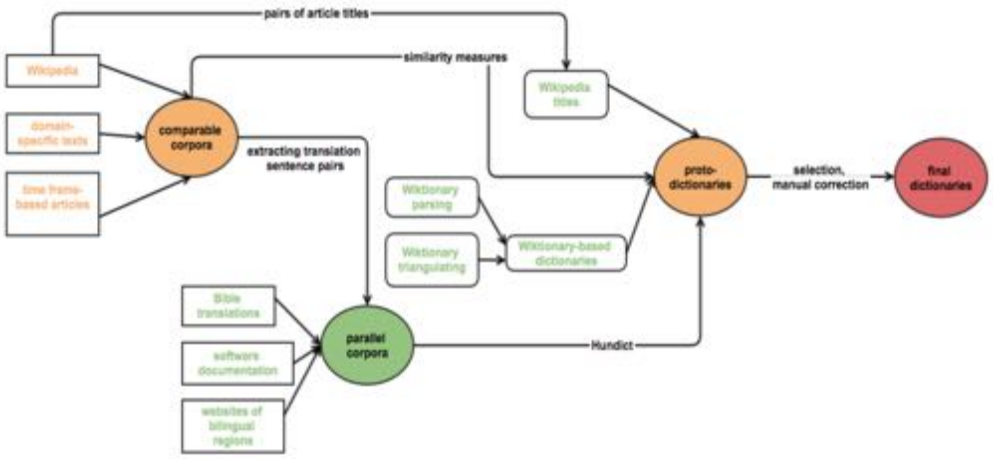


Figure 11: The main steps of the workflow of creating dictionaries.

with this task is the lack of text processing tools for these small languages. We will try to use language-independent tools and methods for extracting as much translation candidates as possible.

Having the proto-dictionaries, they will be merged, then translation candidates with the highest confidence measure will be chosen. As the last step, each entry will be manually checked by native speakers and uploaded to Wiktionary. While the expansion of Wiktionary is targeting the direct support of the mentioned language communities, other materials (dictionaries, corpora, models) will enable the development of additional tools as all of them will be publicly available after the end of the project.

The research reported in the paper is conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #107885.

Notes

[1] `https://github.com/zseder/hundict`.

References

Moseley, C. (eds.).(2010) *Atlas of the World's Languages in Danger. 3rd ed.* Paris, UNESCO Publishing. Online version: `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Simon, E.; Lendvai, P.; Németh, G.; Olaszy, G.; and Vicsi, K. (2012) *A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age*. Georg Rehm and Hans Uszkoreit (Series Editors): META-NET White Paper Series. Springer.

Kornai, A. (2013). Digital Language Death. *PLoS ONE*, 8(10).

Ács, J. (2014) Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

Literary constellations. A Digital Humanities approach to the study of literary salons in Mexico during the 19th century

Silvia Gutiérrez
Würzburg Universität
silviaegt@gmail.com

In this paper I will propose a way of understanding literary history and the establishment of a written culture in Mexico through Geo-spatial and network visualization of literary associations. Following Bourdieu's *rules of art* I will consider salons as "key elements in the social genesis of the literary field" (Bourdieu, 1996, p.XIX). However, rather than considering them as simple mediums for assuring "stage patronage" (Bourdieu, 1996, p.49), I will return to Matthew Arnold's idea that these associations served as literary tribunals (Arnold, 1914, p.39) i.e. schools of style, and in the Mexican-case, schools of specific aesthetic thoughts that would later reign on national literature expressions.

Ultimately, exploring these schools of style through data compiled in the bibliographic works about them (Perales Ojeda, 1957; Sánchez, 1951) means adding a new point of view to the literary historiography which, traditionally, tends to categorize its actors either merely temporarily or aesthetically. Belem Clark de Lara already showed her concern on these divisions and opted to talk about constellations rather than generations to stress the idea that literary groups are not necessarily formed by contemporary authors (in the same way constellations may have stars of different ages), and that their unity is drawn by an imaginary line (Clark de Lara, 2005, p.16).

By identifying these associations' location, duration, and participants I have created a relational database from which I derived some graphic representations that I hope will provide new ways of exploring the coordinates of these networks. The temporal and geographical distribution of these groups will be displayed using DARIAH's Geo-browser (first figure below) while the lines connecting its participants' network, will be sketched with the visualization and network analysis statistic possibilities integrated in Gephi (second figure below). Finally, I will approach these methods critically by addressing two questions: is the visual display of these groups and their participants' relevant for our understanding of them? And if so, how do digital methods help us portray a new way of comprehending them?

References

Arnold, M. (1914) 'The literary influence of Academies', in Francis William Newman (ed.) *Essays by Matthew Arnold: including Essays in criticism, 1865, On translating Homer (with F.W. Newman's reply), and five other essays now for the first time collected*. London; New York: Humphrey Milford : Oxford University Press. pp. 37–63.
Bourdieu, P. (1996) *The Rules of Art: Genesis and Structure of the Literary Field*. California: Stanford University Press.
Clark de Lara, B. (2005) '¿Generaciones o constelaciones?', in Elisa Speckman Guerra & Belem Clark de Lara (eds.) La república de las letras: asomos a la cultura escrita del México decimonónico. Ambientes, asociaciones y grupos: movimientos, temas y géneros literarios. México: UNAM. pp. 11–46.
Perales Ojeda, A. (1957) *Asociaciones literarias mexicanas: siglo XIX*. México: Impr. Universitaria. Sánchez, J. (1951) Academias y sociedades literarias de Mexico. Chapel Hill: University of North Carolina.

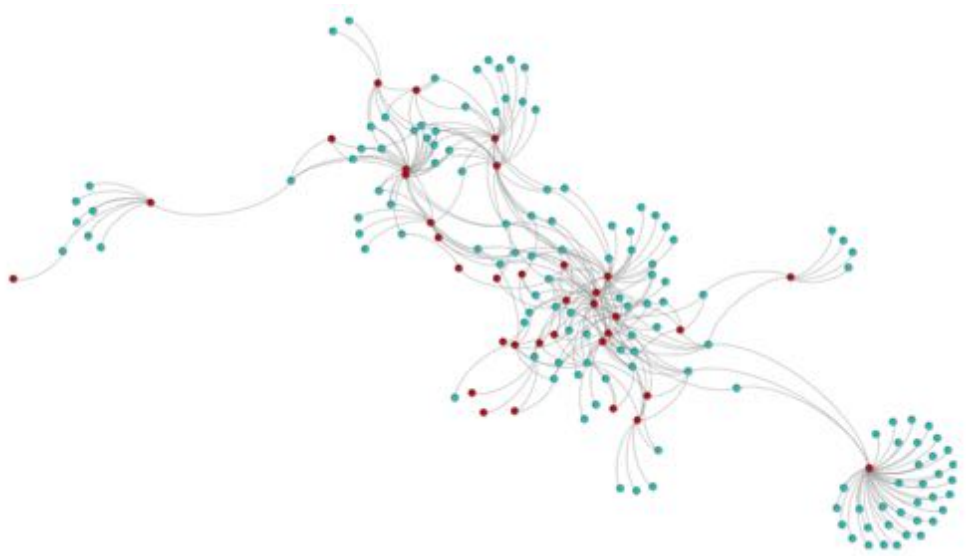


Figure 12: 208 Literary associations time-mapped with DARIAH's Geobrowser.

Figure 13: 36 unique Mexico-City-based associations (red), and 138 different members (blue) plotted with the Yifan Hu Proportional layout algorithm.

Location Extraction Tool

Rosa Merino Claros
University of Amsterdam
r.merinoclaros@uva.nl

Alex Olieman
University of Amsterdam
olieman@uva.nl

Our aim is to build a location extraction tool for text files. The main idea of this engine is the following: to locate within a text file the words that point to a location – a street, a city, a region or a country – and create as output a map with all the locations found in the text. This kind of tool has many possible interesting applications in the digital humanities: for journalists and social science researchers exploring newspapers, for historians working with historical documents, for the analysis of biographies, etc. But the main advantage of such an engine consists of its potential as visualisation tool: at a glance the researcher gains an idea of the spatial interaction described in the text file.

Location extraction encompasses much more than just a lookup of every single word of a text in a gazetteer. The result of the place name lookup – geoparsing – then needs to be disambiguated. The main difficulty is to disambiguate place names by distinguishing places from persons and by selecting the most likely place out of a list of homographic place names.

In order to adapt this tool to the concrete needs of the CREATE project – which offers an interdisciplinary platform for analysing and visualising the making of creative cities like Amsterdam – we have added a semantic filtering feature. In this sense our goal is also the possibility to filter out all mentions of a given city that are related to its cultural activity by spotting names associated with the creative industry e.g. cinemas, painters, theatres, museums, etc., and highlighting the locations where such entities have operated. The final output is a map marking all named locations linked to the text fragment where the creative industry mentions appear.

References

Pouliquen B., Kimler M., Steinberger R., Ignat C., Oellinger T., Fluart F., Zaghouani W., Widiger A., Forslund A., Best C. (2006), 'Geocoding multilingual texts: Recognition, disambiguation and visualisation', In *Proceedings of LREC-2006*.

MAGIS Brugge: a 16th-century bird's-eye view on Bruges as a digital stage for public urban history

Elien Vernackt
Bruggemuseum
elien.vernackt@brugge.be

Bram Vannieuwenhuyze
Caldenberga KU Leuven
bram.vannieuwenhuyze@arts.kuleuven.be

In 1562, the painter Marcus Gerards drew a splendid bird's-eye view on Bruges. The town council requested him to present Bruges more accessible than it actually was. The result was an extremely detailed map that gives a unique view of sixteenth-century Bruges. Historians, art historians and archaeologists know the document for a long time and use it to illustrate their books and articles. It is still very popular amongst the wider public as well. A lot of reproductions adorn living rooms inside and outside Bruges.

The digital *MAGIS Brugge* project aims to go beyond this 'traditional' use of the sixteenth-century map. City maps are always made for a specific purpose: as a means to find the way, as a piece of art that could decorate a place, or, in this case, as a tool for propaganda in order to attract foreign merchants to a decaying port. Yet, hiding the map in a dark corner of an archive has never been the original purpose a cartographer had in mind. The *MAGIS Brugge* project has the ambition to unlock and valorise the enormous mass of information of this sixteenth-century masterpiece to a very diverse audience. In this respect, the project aims to create a new 'stage' for our historical knowledge on Bruges and to stimulate visitors to discover the history of the city.

During the first phase the project offered the opportunity to digitize the bird's-eye view with GIS- technology and to build a scientific knowledge platform accessible for scholars. During the second phase, the project partners aimed to open up the information to the wider audience by creating a user-friendly website. It was necessary to work out clear instructions about data input in order to create a coherent database. In addition we had to find out what the diverse target groups actually hope to discover on a website with a digitized historical map. This might appear rather simple, but in fact we faced several complications and had to make compromises. In this paper we want to discuss some of the difficulties that we have encountered and the solutions that were chosen for the *MAGIS Brugge* project. Hence we hope to share experiences and inspire scholars, museums, city councils, associations and everyone interested to see the potential of digitizing and unlocking historical maps online.

Figure 14: Printscreen of the user-friendly website at `www.kaartenhuisbrugge.be/magis`.

Mapping Digital Humanities projects

Stef Scagliola
Erasmus University Rotterdam
scagliola@eshcc.eur.nl

Barbara Safradin
Erasmus University Rotterdam
b.safradin@gmail.com

Almila Akdag
eHumanities group, KNAW
alelma@gmail.com

Sally Wyatt
eHumanities group, KNAW
sally.wyatt@ehumanities.knaw.nl

Andrea Scharnhorst
DANS / eHumanities KNAW
andrea.scharnhorst@dans.knaw.nl

Digital Humanities (DH) has grown, and now exhibits many of the characteristics of an emergent field, moving beyond the early stage populated by lonely pioneers and early adopters (Whitley, 1984/2000). The field has its professional organizations, local and international conference series, dedicated journals, specific funding programmes, and an increase in DH minors, master and PhD programmes. DH also regularly features in bibliometric mapping exercises, however formal scholarly communication covers only a part of DH. Digital Humanities expresses features of a virtual community and engages with mundane and specific infrastructures, platforms, tools and software – from Twitter to the ePistolarium, to name two examples. Missing consolidated data about DH activities is one reason, why it is not easy to monitor and map DH. The still growing community DH with new centers, initiatives, projects almost emerging daily, has a substantial need for registries, catalogues, interactive maps to resources, in short for intelligent information management. This paper responds to those needs. It builds on a course registry (https://dariah.uni-koeln.de/), initiated among others by some of the authors of this paper, and currently supported by DARIAH and CLARIAH. As a follow up, we present in this paper a project registry. We present the structure of the database, ways of data collection and cleaning, and possible user interfaces. Ordering and classifying projects in DH is another aspect we discuss. We argue that such a database will:

1. Support researchers in finding related projects and extent their networks
2. Deliver empirical evidence for funding bodies to shape future policies in this area
3. Support the 'building on other experiences' in the field of DH and thus contribute to its consolidation. In this paper we present a prototype, and first data collection focusing on The Netherlands. We discuss possible interactive interfaces for browsing and searching. We also discuss issue of sustainability and maintenance. What are short-, mid- and long-term functions of such a project registry? How do we envision domain specific information management to be integrated in generic and general achievements in Research Information Systems (this is where project information belongs to)? What is the relationship between such a registry and

systems as NARCIS or VIVOweb? How can we in the design of the database achieve a high degree of potential interoperability?

References

CLARIAH – Common Lab Research Infrastructure for the Arts and Humanities, www.clariah.nl.

Whitley, R.J. (1984/2000) *The Intellectual and Social Organization of the Sciences* (Second Edition). Oxford: Oxford University Press.

Mapping urban multilingualism through Twitter

Enrique Manjavacas
Freie Universität Berlin
enrique.manjavacas@gmail.com

Ben Verhoeven
University of Antwerp
ben.verhoeven@uantwerpen.be

With the advent of globalisation, the interweaving of different languages in urban environments has acquired a high degree of complexity and has become an attractive topic for sociolinguistic research in recent years. Conveniently, in 2009 the microblogging service Twitter added an option that allowed users to attach geolocation information to their broadcasted tweets. Although only 1,6% (Leetaru et al. 2012) of the Twitter stream is actually shipped with geolocation information, these geo-referenced tweets can be utilised to address general issues of urban multilingualism that otherwise elude research due to the high cost of data collection.

In this project, we test, as a proof of concept, the viability of utilising Twitter data to analyse large-scale distributional patterns of language use in four European cities (Amsterdam, Antwerp, Berlin and Brussels, alphabetical order). We use our growing database of tweets, geolocated to the cities of interest, as collected and filtered from the Twitter stream. We experiment with a majority vote approach based on preexistent language identification system in order to guarantee precise language identification, given that this task is an essential preparatory step of the project. Moreover, different preprocessing steps are carried out and evaluated in order to account for potentially distorted data - filtering out bots, identifying tweets by tourists etc.

Finally, we devise and test various techniques for aggregating and visualising geolocated data and evaluate the data against known demographic figures (official censuses and open source datasets on population etc.). The resulting dataset is suitable for addressing questions such as the distribution of languages across urban areas, spotting touristic areas, analysing differences of languages across day/night population patterns, etc. As a result, we aim to obtain insight into the urban multilingual picture and simultaneously into the bias that Twitter data may introduce therein.

Measuring Impact. Using Topic Modeling to analyse the influence of the Second Vatican Council on Dutch public debate

Maarten van den Bos
Utrecht University
m.j.a.vandenbos@uu.nl

How to measure the influence of ideas on broader public debate is a question that since long has puzzled historians. In their recent volume on intellectual history, Samuel Moyn and Andrew Sartori rightfully stress the importance of impact in the historiography of ideas, but fail to provide practical guidelines that enable researchers to meet their standards. (Moyn and Sartori 2013) Potentially, the large scale digitation of sources that reflect public discourse such as newspapers and periodicals could help, although the road ahead is dark and full of terrors (Bingham 2010; Broersma 2012).

Since ideas seldom come in the form of a set of searchable key-words, I propose a data driven approach to mine large digitized corpora for impact and influence of key texts. In my own work, on the impact and reception of the Second Vatican Council, I have tested several possible tools and techniques. Key question in the historiography of the council is how influential the documents produced by the council really were. (van den Bos 2014; Faggioli 2012) Building upon the work of others, I have come up with a method to make more empirically sound claims with regard to this question. The method contains three consecutive steps:

- Topic modeling the conciliar documents [1]
- Use the topics as large OR queries to search the digitized newspaper collection of the Dutch National Library [2]
- Select and analyse texts that contain a minimum 8 out of 10 words produced by step one.

The first outcomes are promising. To give but one example: topic modeling the pastoral constitution about the role of the church in the modern world, *Gaudium et Spes*, produces among others the string of words 'humans, world, all, love, Christ, time, spirit, make, dignity, calling' [3]. Using this string of words as a large OR query gives back more than four million results. I have selected and analysed the first hundred articles, that all contained a minimum of eight out of ten words. All articles selected were relevant, roughly half was about the debate on 'Scheme XIII' as the text was called before ratified by the council on December 7 1965. The other half was on themes like development aid, catholic politics and war and peace, discussions suspected in historiography to be influenced by the Second Vatican Council. (Fergusson 2004, 130; Van den Bos 2015) By repeating these steps for all sixteen conciliar documents, a palette of different themes and discussions can be mapped that was influenced by conciliar ideas and vocabulary.

In my paper I would like to not only further elaborate on the proposed method to analyse the impact of new ideas on public debate, but also would like to show how

the method enables me to answer an important historical question on the impact of the Second Vatican Council on Dutch public debate.

Notes

[1] For Topic Modeling I use the Mallet GUI (`https://code.google.com/p/topic-modeling-tool`), an open source graphical user interface tool for Latent Dirichlet Allocation topic modeling. Settings: 10 topics of 10 words; 200 iterations; 0.1 topic proportion threshold.

[2] For searches in the collection of the Dutch National Library I use Texcavator, a program developed by researchers at the universities of Utrecht and Amsterdam. (van Eijnatten, Pieters and Verheul 2014).

[3] I have used the Dutch translation of *Gaudium et Spes*. Original string of words generated by Mallet: 'mensen, wereld, alle, liefde, christus, tijd, geest, maken, waardigheid, roeping'.

References

Bingham, A., 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians', *Twentieth Century British History* 21 (2010) 225-231.

Bos, Maarten van den, 'Een nieuwe bijdrage aan de receptiegeschiedenis van het Tweede Vaticaans Concilie', *Religie en Samenleving* 9 (2014) 211-234.

Bos, Maarten van den, *Mensen van goede wil. Pax Christi*, 1948-2013 (Amsterdam 2015).

Broersma, M., 'Nooit meer bladeren? Digitale krantenarchieven als bron', *Tijdschrift voor Mediageschiedenis* 14 (2012) 29-55.

Eijnatten, J. van, T. Pieters and J. Verheul, 'TS Tools: Using Texcavator to map public discourse', *Tijdschrift voor Tijdschriftstudies* 35 (2014) 59-65.

Fergusson, D., *Church, State and Civil Society* (Cambridge 2004).

Moyn, S. and A. Sartori, 'Approaches to Global Intellectual History', in: Id. (eds.), *Global Intellectual History* (New York 2013) 3-33.

Measuring the Use of Collections Before and After Publication in Wikimedia

Trilce Navarrete
University of Southern Denmark
trilce.navarrete@gmail.com

The mission of museums worldwide revolves around giving access to humans' knowledge. Intergenerational transfer of knowledge about collections is taking an important new dimension: with Wikipedia, people all over the world can have access to potentially everything housed and managed by the memory institutions. But, what is actually the impact of publishing online? Based on two Dutch museum cases, access to collections will be analyzed comparing publication of analogue collections with publication online using the Wikimedia BaGLAMa2 tool. It will be argued that publication through Wikimedia substantially increases access to collections, particularly of the lesser-known collections.

Digitization has represented not only a chosen technical option for the management of information but has also fundamentally changed the way to access and (re)use information in all areas of life. However, limited data is available to understand the actual impact of emerging knowledge transfer platforms, including Wikipedia. Most quantitative analysis related to digital heritage collections derives from the library science tradition, particularly the analysis of use and seeking behaviour in order to understand user needs and improve the information service in museums (Skov and Ingewersen, 2008), archives (Yakel, 2004; Zhang and Kamps, 2010) or libraries (Connaway, 2011). Results show that users can have different information needs at different times and that convenience of access remains an important factor when accessing content. Users further favour full access to content, including images, sounds or full text rather than only a bibliographic reference. Allocation of resources would benefit from a better understanding of users' preferences and information needs. In the Netherlands, only a handful of museums have started collaborations with Wikimedia. This contribution will review the history of the Tropenmuseum, ethnographic museum in Amsterdam and of the Rijksmuseum, art and history museum in Amsterdam, and their approach to open data and their collaboration with Wikimedia. Analysis will include a follow-up of the access to objects through the archive (in the analogue world) and through the BaGLAMa2 tool (in the Wikipedia world). The presentation will consider issues of collection origin and origin of users, GLAMs as illustration, organizational process when publishing the collections in Wikimedia and expectations of success. Particular attention will be given to the democratization of access to collections: all objects, from any part of the world, can be potentially positioned in any context to be accessed from all over the world. The repositioning of digital objects, from the physical museum and the institutional website into an open platform supporting collaboration and exchange, can potentially change the shape of the long tail.

References

Histoire de la culture des plantes sucrières

L'**histoire de la culture des plantes sucrières** couvre une période allant de la haute Antiquité à nos jours.

Mais c'est surtout à partir du milieu du XVIIe siècle, avec le développement du mercantilisme et du colonialisme, que commence la période dite industrielle avec la Caraïbe qui devient la principale région mondiale pour la production du sucre obtenu à partir de la canne.

Cette période connaît une expansion à marche forcée au siècle suivant, marquée par le commerce triangulaire vers les îles françaises et anglaises. Elle n'a cédé cette place qu'à la fin du XIXe siècle avec l'abolition de l'esclavage et le

Planteur de canne à sucre indonésien exhibant son outillage; époque : Empire colonial néerlandais.
Crédit photo : Tropenmuseum.

Figure 15: Screenshot of a negative from the Tropenmuseum made available in Wikimedia and used in a Wikipedia article. Source: `http://fr.wikipedia.org/wiki/Histoire_de_la_culture_des_plantes_sucrières`.

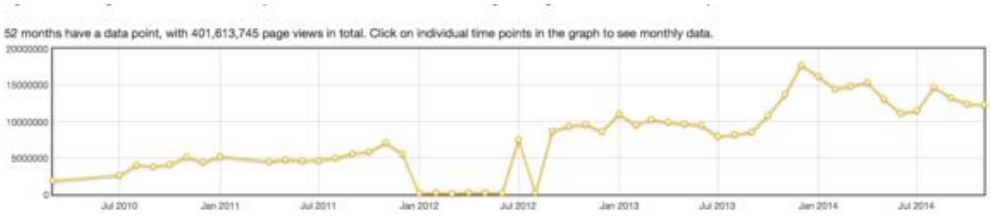


Figure 16: Page views of Wikipedia articles containing images from the Tropenmuseum. Source: `http://tools.wmflabs.org/glamtools/baglama2`.

Jancic, Maja Bogataj, et al. (2015) LAPSI Policy Recommendation N.5 The Proposed Inclusion of Cultural and Research Institutions in the Scope of PSI Directive. Brussels: EC.
Connaway, Lynn, et al. (2011) "'If it is too inconvenient I'm not going after it:" Convenience as a critical factor in information-seeking behaviours' in *Library & Information Science Research*. 33(3):179-190. Navarrete, Trilce (2014) *A History of Digitization: Dutch Museums*. PhD dissertation. University of Amsterdam.
Skov, Mette and Pter Ingwersen (2008) 'Exploring information seeking behaviour in a digital museum context' in *IIiX '08 Proceedings of the second international symposium on Information interaction in context*. New York: ACM, pp. 110-115.
Yakel, Elizabeth (2004) 'Seeking information, seeking connections, seeking meaning: genealogists and family historians' in *Information Research*. 10(1). Paper 205.
Zhang, Junte and Jaap Kamps (2010) 'A Search Log-Based Approach to Evaluation' in M. Lalmas et al. (eds.), *Research and Advanced Technology for Digital Libraries*, LNCS 6273, pp. 248-260. Berlin: Springer.

Modelling Discussion Topics to Improve News Article Tagging

Chris Emmery
University of Antwerp
chris.emmery@uantwerpen.be

Menno van Zaanen
Tilburg University
mvzaanen@uvt.nl

Online news articles are often labelled by their writers with a set of tags to topically frame their content. For many news websites, the tags allow for easier retrieval of articles by news readers after the articles have vanished from the front pages. In addition, tags cluster articles by topic, granting the ability to quickly search through multiple articles on the same topic.

We show that human-provided tags have two major shortcomings. Firstly, they often result in several uninformative tags which are seldom reused. Secondly, the social context of the article is not always fully described by the tags. Readers may have associations or ideas in relation to an article that were not predicted by its writer and, hence, corresponding tags are missing. The first problem can be resolved by frequency filtering of the tags. We argue here that the second problem can be resolved by detecting the latent topics in online discussions that directly relate to the article (for instance, discussions in the comment section below news articles) and link them to the tags of the article that were provided by the writer. The intuition behind this idea is that information taken from the context of previously written articles can be leveraged to improve the quality of the tags. Additionally, information from the previously identified latent topics can be reused to assign tags to new articles.

In this research, we employ a supervised topic model to a collection of news articles and their associated comments to detect latent topics. By utilizing the tags that are already assigned to the articles, we are able to train a Labeled Latent Dirichlet Allocation (L-LDA) model in a supervised setting. By doing so, we can evaluate tags proposed by the fitted model for unseen articles. This is done in a ranked setting using Mean Average Precision (MAP) against the original human-provided tags. We will demonstrate the influence of the social context of a news article, taken from the related discussion, on the performance of the overall model. Additionally, we can investigate the quality of the model by comparing its inferred individual tags to those of the writer. Finally, this approach allows for a close collaboration between the system and the writer of news articles to improve tag assignment, with the aim of improved searches for related articles.

Monitoring Online User Behaviour. The case of the Newstracker

Martijn Kleppe
Vrije Universiteit
m.kleppe@vu.nl

Irene Costera Meijer
Vrije Universiteit
Icostera.meijer@vu.nl

Digitalization of journalism enables news organizations to monitor the behavior of online news users by using metric tools such as Google Analytics. Journalists and news executives take these data as good measurements of audiences' interest in news, but they cover only their own websites. To create a full picture of all online news consumption, research companies create lists of aggregated visits to websites. However, this type of research does not consider website visitors as citizens but as commodity (Richardson, 2007, p. 79; Usher, 2013), giving advertisers detailed information on how to reach their target audience in the most efficient manner. Feedback to journalists is limited to presenting the most clicked items, often leading to the critique that most news users are more interested in trivial news than in public affairs (Boczkowski, Mitchelstein, & Walter, 2011; Karlsson & Clerwall, 2013; Nederlandse Nieuwsmonitor, 2013). Relatively little is known about the content of the visited news articles, let alone the everyday patterns of individual news consumption (Costera Meijer & Groot Kormelink, 2014).

Therefore, this paper will present the set-up and first results of the research tool 'Newstracker' developed especially for monitoring news consumption from the angle of the user. We first installed a proxy on the laptop of a group of 50 respondents who use this device regularly for the consumption of news and information. This set-up is in line with Findahl (2013) who investigated the online behaviour of an American family and Menchen-Trevino (2012) that used a special-designed proxy to monitor the exposure to political communication during the November 2010 U.S. general election campaign. However, these academic studies only report the website titles that have been visited. Our set-up goes two steps further. We do not only monitor the website titles but also the actual visited URLs and we crawl all textual and visual contents of the visited websites. Since one of the problems when monitoring a person's online behaviour is the magnitude of the data that is being collected (Batista & Silva, 2002; Manovich, 2012; Vicente-Marino, 2013, p. 43), we deploy automated content analyses techniques (Atteveldt, 2008; Bhulai, Kampstra, Kooiman, Koole, & Kok, 2012) to detect the topics that are being discussed in the news items. This enables us to calculate the topical online news consumption during the day.

In this paper we will discuss the set-up of our research tool and its applicability for other types of Digital Humanities research such as user studies focussing on formulating requirements, based on existing user behaviour.

References

Atteveldt, W. van, 2008. *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. BookSurge.

Batista, P., Silva, M., 2002. Mining Web Access Logs of an On-line Newspape. *Proceedings of the Workshop on Recommendation and Personalization in eCommerce of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, Malaga.

Bhulai, S., Kampstra, P., Kooiman, L., Koole, G., Kok, B., 2012. Trend visualization on Twitter: What's hot and what's not In: Data Analytics. *IARIA*, Barcelona, pp. 43–48.
Boczkowski, P.J., Mitchelstein, E., Walter, M., 2011. Convergence Across Divergence: Understanding the Gap in the Online News Choices of Journalists and Consumers in Western Europe and Latin America. *Commun. Res.* 38, 376–396. doi:10.1177/0093650210384989.
Costera Meijer, I., Groot Kormelink, T., 2014. Checking, Sharing, Clicking and Linking. *Digit. Journal*, 1–16. doi:10.1080/21670811.2014.937149.
Findahl, O., Lagerstedt, C., Aurelius, A., 2013. Triangulation as a way to validate and deepen the knowledge about user behavior. A comparison between questionnaires, diaries and traffic measurement, in: *Audience Research Methodologies: Between Innovation and Consolidation*. New York, pp. 54))69.
Karlsson, M., Clerwall, C., 2013. Negotiating Professional News Judgment and "Clicks." *Nord. Rev.* 34, 65–76. doi:10.2478/nor-2013-0054.
Kleinneijenhuis, J., Atteveldt, W. van, 2006. Geautomatiseerde inhoudsanalyse, met de berichtgeving over het EU- referendum als voorbeeld, in: *Inhoudsanalyse: Theorie En Praktijk*. Kluwer, pp. 227–250.
Manovich, L., 2012. *How to Follow Software Users?*
Menchen-Trevino, E., Karr, C., 2012. Researching Real-World Web Use with Roxy: Collecting Observational Web Data with Informed Consent. *J. Inf. Technol. Polit.* 9, 254–268. doi:10.1080/19331681.2012.664966.
Nederlandse Nieuwsmonitor, 2013. 'Seksmoord op horrorvakantie: de invloed van bezoekersgedrag op krantenwebsites op de nieuwsselectie van dagbladen en hun websites'.
Richardson, J.E., 2007. *Analysing newspapers: an approach from critical discourse analysis*. Palgrave Macmillan, Basingstoke [England]; New York.
Usher, N., 2013. Al Jazeera English Online. Understanding Web metrics and news production when a quantified audience is not a commodified audience. *Digit. Journal* 1, 335–351. doi:10.1080/21670811.2013.801690.
Vicente-Marino, M., 2013. Audience research methods. Facing the challenges of transforming audiences, in: *Audience Research Methodologies: Between Innovation and Consolidation*. New York, pp. 37–53.

On Seeking the Other: An Outlook on Digital Doppelgänger Trends

Alia Soliman
University College London
soliman.alia@gmail.com

The *doppelgänger* is a term which crystallized in the 18th and 19th century and was widely associated with the onset of doom. This belief was reflected and reinforced in literary creations such as *Peter Schlemihl* by von Chamisso (1814), *William Wilson* by Poe (1939), and the film *The Student of Prague* (1913) in which the self, and often the copy, is destroyed after it meets its counter self. While we can still see remnants of this negative conception, the association between meeting our doubles and harm, madness, or death is arguably changing in the digital age.

With the introduction of the digital era, our image consumption has risen significantly. The ease of taking pictures using digital cameras and smart phones, and the equal ease of posting them on the Net and social media websites has changed the nature of the image, in terms of constitution, accessibility, and distribution. On the Internet, the phenomenon centering on finding lookalikes or digital *doppelgänger* is increasingly on the rise, marking a heightened sense of self-awareness, and a shift in its nature. The *doppelgänger* has undergone a change in both its nature (an elusive symbolic concept visualized in a more tangible and photographic form) as well as in the willing acceptance of seeing the self and its lookalike in the same moment, to the extent of even being photographed together [2]. In fact, digital imagery has inverted the anatomy of the *doppelgänger* figure: an image on the Internet gives the impression of being there forever, and thus serves as a preservation against death and oblivion rather than being their trigger, a reversal of classical conception.

I will analyze two digital *doppelgänger* campaigns through contemporary theories of cultural psychology that shed light on changes to the self and the *doppelgänger* figure. An application of Kenneth Gergen's Saturated Self concept and Hubert Hermans's Dialogical Self Theory will show how the marriage between the *doppelgänger* figure and the digitized image celebrates multidimensionality and self-authorship, further, that the new world of modernity, globalization, and digital opportunities impact our consciousness which has evolved to become more social, discursive, and reliant on the other. My first example features a young woman's ardent search for her *doppelgänger* through social media websites. In 2011, Sophie Robehmed, a British-Lebanese journalist began looking for her own *doppelgänger* [3]. The second example involves a photographer's project to allocate pairs of lookalikes and snap them in photos together. In 2006 François Brunelle launched a project called "I Am Not a Lookalike". Brunelle's campaign produced 140 pairs of successful matches from around the globe [4].

In digital *doppelgänger* trends, the culture of the copied self arises from the pursuit of self-innovation and new positions for the self. My paper will attempt to explain the way the self adjusts to new types of experiences and to the multiplicity of positions afforded to it by the expansion of our world. Relevant postmodern theories suggest that

the self is a socially constructed entity that changes its constitution in order to adapt to its dynamic surrounding. The increasing fascination with the culture of lookalikes [5] further suggests the need for alterity as well as the constant reevaluation of self-positions. New positions as well as new modes of thought emerge to allow previously banned or feared imagery to surface and be part of a shifting self-image. I will show through my research that new *doppelgänger* trends treat the self as a hybrid entity that contains, not one, but several selves.

Notes

[1] An article is forthcoming in "The International Journal of the Image", see `www.ontheimage.com`.

[2] An increase in visual presentations of the double self is exemplified in three films released in 2013, *The Double*, *Enemy*, and *The Face of Love*.

[3] Robehmed's video entitled "Please world, help me find my doppelgänger' is available on Youtube (`http://www.youtube.com/watch?v=u87LYolDrmY`)

[4] You can see these images on `http://www.francoisbrunelle.com/index.php?id=3&lang=En`.

[5] The examples given here are but a few in the contemporary fascination with the culture of lookalikes. Websites such as `http://www.ilooklikeyou.com/`, `http://www.findmydoppelganger.com/`, and `http://www.reddit.com/r/Doppleganger/` help individuals to find their *doppelgänger*. More structured and professional projects include Martin Schoeller's photography campaign which investigates why identical twins differ in features and tastes later in life. Spanish photographer María Zarazúa has a project titled 'Parte de ti', which in Spanish translates as 'Part of you'. Zarazúa's images investigate the similarities and differences between identical twins through digital photography. The variation on this trend is endless and gaining momentum.

References

Ardent, Paul. "I Am Not a Lookalike". *The Guardian* (January 12, 2006): 23.

Belting, Hans. (Winter 2005). "Image, Medium, Body: A New Approach to Iconology." *Critical Inquiry* 31.2: 302–19.

Freud, Sigmund. 2003. *The Uncanny*. Trans. David Mclintock. London: Penguin Books. Gergen, Kenneth. 2000. The Saturated Self. New York: Basic Books.

Hermans, Hubert. 1999. "Dialogical Thinking and Self-Innovation". *Culture & Psychology* 5: 67.

Hermans, Hubert. 2012. "Dialogical Self Theory and the Increasing Multiplicity of I-Positions in a Globalizing Society: An Introduction." Ed. Hubert Hermans. *Applications of Dialogical Self Theory: New Directions for Child and Adolescent Development* 137: 1–21.

—. 2001. "The Dialogical Self: Toward a Theory of Personal and Cultural Positioning." *Culture & Psychology* 7: 243–281.

—. 2004. Introduction: "The Dialogical Self in a Global and Digital Age". *Identity: An International Journal of Theory and Research* 4: 297–320.

Hermans, Hubert and Giancarlo Dimaggio. 2007. "Self, Identity, and Globalization in Times of Uncertainty: A Dialogical Analysis". *Review of General Psychology* 11.1: 31–61.

Hermans, Hubert, Kempen, H. J. G., & Van Loon, R. J. P. 1992. "The Dialogical Self: Beyond Individualism and Rationalism". *American Psychologist* 47: 23–33.

Holoquist, Michael. 1990. *Dialogism: Bakhtin and his World*. London: Routledge.

James, Henry. 2005. *The Jolly Corner*. Kindle File.

Macías, Javier and Rafael Núñez. December 2011. "The Other Self: Psychopathology and Literature". *Journal of Medical Humanities* 32.4: 257–267.

Markus, Hazel, and Paul Nurius. 1986. "Possible Selves." *American Psychologist* 41: 954–69.

Nebenzahl, Donna. (26 February, 2006). "Buddy Double". *The Gazette*: A18.

Rank, Otto. 2009. *The Double: A Psychoanalytic Study*. Trans. Harry Tucker Jr. Chapel Hill: U of North Carolina P Enduring Editions.

Robehmed, Sophie. March 6, 2013. "Doppelgänger: Desperately Seeking My Lookalike". *BBC News: Technology* (*http://www.bbc.com/news/technology-21383109*).

Salgado, João, Carla Cunha, and Tiago Bento. 2013. "Salgado Positioning Microanalysis: Studying the Self Through the Exploration of Dialogical Processes". *Integr Psych Behav* 47: 325–353.

Schmid, Astrid. 1996. *The Fear of the Other: Approaches to English Stories of the Double (1764-1910)*. Bern and New York: Peter Lang.

Underwood, Mitya. July 5, 2013. "Seeing Double: a Dubai Expat's Hunt for her Doppelganger" (`http://www.thenational.ae/lifestyle/seeing-double-a-dubai-expats-hunt-for-her-doppelganger#ixzz31nPXyWUa`.

People on the Move and Cultural Heritage in a Digital Era. Participative and biographical collecting at the Red Star Line Museum Antwerp

Marie-Charlotte Le Bailly
Red Star Line Museum Antwerp
marie-charlotte.lebailly@stad.antwerpen.be

The Red Star Line Museum not only explains the process of the historical emigration to America via Antwerp to a wide and varied audience, but also makes the connection with actual migration to Western Europe and in particular Belgium. Migration is a universal phenomenon. We are all, in some way or another, the product of our own migration movements or those of our parents/forebears. The reconstruction of personal migration stories is a good means of shedding light on all aspects of this fascinating subject. Migration stories or migrants' biographies, both past and present, therefore play an important role in the scenography and educational department, as well as in the presentation of our digital collections.

Over the past few years, the Red Star Line Museum has collected about 1500 personal stories of migration, today and in the past, through research of staff-members and participation of the public. The museum actively invites visitors to participate and share their own migration stories or that of their relatives through different means. The resulting biographical collection comprises a wide range of personal stories of migration, related objects and documents, and contextual information needed to understand these. We make these available to our visitors through our digital 'Warehouse' (*'t Magazijn*). *'t Magazijn* is not only our content management system, but it has also been designed as tool to manage our (digital) collection of stories of migration.

In this paper I will discuss some of the challenges we meet in the management, use and accessibility of our biographical (data) collections, i.e. finding the right balance between our own needs and resources and meeting our public responsibilities/duties as a museum. I will also address the following topics: biographies and linked data, dealing with data in heterogeneous datasets, the use of taxonomies and thesauri, and finally standards and best practices for the encoding and processing of (biographical) data and collection management (such as SPECTRUM Museum Collection Management Standard and international standards for the exchange of biographical data like A2A (Archive 2 Archive), BioDes and GEDCOM). It is fair to admit that we do not have all the answers and that we still struggle with some issues regarding conceptualization, public participation, crowd sourcing and reliability of the data. Being so varied in terms of content, type and sources, the collection of personal stories poses many methodological challenges with regard to categorization, presentation and scientific appreciation. In the related demonstration I will show the public interface to the digital Warehouse (*'t Magazijn*) as well as the CMS we developed to manage our collection of personal stories.

References

Figure 17: Collection Management in *'t Magazijn*.

'Art and Architecture Thesaurus (AAT)', *Kennisbank DEN Kenniscentrum Digitaal Erfgoed* (www.den.nl/standaard/20/).
'A2A model (A2A)', *Kennisbank DEN Kenniscentrum Digitaal Erfgoed* (www.den.nl/standaard/386/).
'BioDes', *Biografisch Portaal van Nederland* (www.biografischportaal.nl/about/biodes).
'BioDes (BioDes)', *Kennisbank DEN Kenniscentrum Digitaal Erfgoed* (www.den.nl/standaard/259/BioDes).
'Spectrum', *Kennisbank DEN Kenniscentrum Digitaal Erfgoed* (www.den.nl/artikel/bericht/3165/).
'Union List of Artists Names (ULAN)', *Kennisbank DEN Kenniscentrum Digitaal Erfgoed* (www.den.nl/standaard/18/).
Babazia, N., e.a. (2014) 'Levensverhalen en storytelling in het Red Star Line Museum', *FARO | Tijdschrift over cultureel erfgoed*, vol. 7: 4, pp. 24-31.
Coret, B. (2013) 'A2A open: goed nieuws voor archiefinstellingen!', Blog ... over het raakvlak van Internet en genealogie door Bob Coret, 08.03.2013 (blog.coret.org/2013/03/a2a-open-goed-nieuws-voor.html).
Coret, B., 'Wat is GEDCOM?', *vraag.en.antwoord* (vraag.en.antwoord.coret.org/entry/33/).
Ernst, M. (2014) 'De som van individuele herinneringen. Van oral history naar verhalen over erfgoed en migratie', *FARO | Tijdschrift over cultureel erfgoed* vol. 7: 4, pp. 16-21, available at www.faronet.be/files/bijlagen/pagina/decembernr_2014_ernst.pdf.
Grever, M. and Boxtel, C. van (2014) *Verlangen naar tastbaar verleden. Erfgoed, onderwijs en historisch besef*. Hilversum: Verloren.
Meijer, M. (2013) *Wie Was Wie. Toelichting A2A Datamodel, versie 1.8 (11.02.2013)* (https://www.dropbox.com/s/pd88vk6kkvg8f8n/A2ABeschrijving_v1.8.pdf?dl=1).
Mormon Migration (mormonmigration.lib.byu.edu/)

SPECTRUM 4.0: The UK Museum Collections Management Standard, available at www.collectionstrust.org.uk/spectrum/the-spectrum-standard.
SPECTRUM Digital Asset Management, available at www.collectionstrust.org.uk/spectrum/spectrum-digital-asset-management.
SPECTRUM-N. Standaard voor collectiemanagement in musea. Versie 1.0, available at www.faronet.be/download-spectrum-n-belgie.
Terminology Management Platform, the preliminary results are shown at athenaplus.thesaurus.condillac.org/.
Unified Thesaurus Haalbaarheidsstudie. Eindrapport. Published on 31.10.2014, available at viaa.be/assets/files/page/downloads/Eindrapport_Unified_Thesaurus_haalbaarheidsstudie.pdf.

Polemics Visualized: Morphological Analysis for Syriac

Hannes Vlaardingerbroek
Vrije Universiteit Amsterdam
hannes@vlaardingerbroek.nl

Marieke van Erp
Vrije Universiteit Amsterdam
marieke.van.erp@vu.nl

Wido van Peursen
Vrije Universiteit Amsterdam
w.t.van.peursen@vu.nl

Syriac, a language from the Aramaic family, has been the lingua franca of the Middle East for centuries. Many important theological documents from the period of the formation of the early church have been written in Syriac. These texts form a considerably large corpus, the published works of Ephrem the Syrian for example already exceed 500,000 words. The theological study of textual corpora of such size would benefit greatly from computational analysis of these texts.

The purpose of the Polemics Visualized pilot project is to explore the possibilities of computational linguistics and natural language processing for use in theological research of Classical Syriac texts. More specifically, we would like to answer the question whether Ephrem the Syrian, who wrote extensive polemics against Bardaisan, a theologian living two centuries earlier, was indeed discussing the same issues as Bardaisan addressed in his only remaining work.

So far, we have successfully trained a tokenizer using Apache OpenNLP and the annotated Syriac resources available at the Eep Talstra Center for Bible and Computer. With the resulting model we succeeded in recognizing 96% of word boundaries in the test data (Vlaardingerbroek, Van Erp, and van Peursen 2015). We then used our tokenizer on the unannotated text of Ephrem, and with the resulting data and that of Bardaisan's annotated work we trained an LDA topic model. The resulting topic model yields some sensible topic-document relations, but not sufficiently useful to aid in finding answers to questions such as the example mentioned above. We now aim to improve the results by morphological and part of speech tagging of the data, which would allow more efficient filtering of the input data for the topic analysis algorithm.

Since the development of a new part of speech tagging algorithm is beyond the scope of the current project, we are now looking into the possibilities to adapt solutions from existing projects to our implementation. NLP software has been developed for other Semitic languages, such as Hebrew and Arabic. However, these approaches rely on large contemporary annotated textual corpora, which are not available for Syriac (Zitouni 2014: 52-55). The only other Syriac NLP software project so far, Syromorph, aims to facilitate the annotation of a large corpus of Classical Syriac, with morphological annotation, links to dictionary entries, and morphological attributes, using a joint pipeline model (McClanahan et. al. 2010). We are now working on the adaptation of the joint pipeline model of Syromorph for morphological analysis and part of speech tagging, in order to improve the results of the tokenizer model and topic analysis.

References

Hannes Vlaardingerbroek, Marieke van Erp, Wido van Peursen (2015) Polemics Visualised: Experiments in Syriac text comparison. *Computational Linguistics in the Netherlands*, Antwerp, February 2015.
Zitouni, Imed. (2014) *Natural language processing of Semitic languages*, Heidelberg: Springer.
McClanahan, P., Busby, G., Haertel, R., Heal, K., Lonsdale, K., Seppi, K., and Ringger, E. (2010) 'A probabilistic morphological analyzer for Syriac'. In: *Proceedings of the 2010 conference on empirical methods in natural language processing*, Massachusetts: MIT pp. 810-820.

Reporting the Empire: The branding of Metropolises and Empire in the Pall Mall Gazette 1870–1900

Tessa Hauswedell
University College London
t.hauswedell@ucl.ac.uk

Melvin Wevers
Utrecht University
m.j.h.f.wevers@uu.nl

The flourishing of the European imperial metropolises, followed by the subsequent demise of European empires and concomitant rise of the American empire in the late 19th century has received ample attention in academic literature(Hall and Rose, 2006; Körner and Smith, 2012; Kroes and Rydell, 2005; Maier, 2007). However, the authors based these processes often on anecdotal evidence or by using quantitative data to study underlying economic processes. In this paper, we aim to turn towards newspapers to study continuities and discontinuities in ideas of empire and metropolises over a longer period [1]. This echoes historians David Armitage's and Jo Guldi's recent plea (Guldi and Armitage, 2014) for a return to the analysis of longer-term narratives at a time when cultural historians have shown an increasing focus on micro-histories.

With the availability of large-scale digitized sources and text mining techniques, the study of long-term cultural processes is now possible in ways hitherto impossible. For this paper, we use the digitized newspaper the Pall Mall Gazette, which has been fully digitized with high-quality OCR recognition for the period between 1870 and 1900 [2]. The Pall Mall Gazette was a London-based newspaper, which upon its foundation in 1865 displayed a conservative outlook. In 1883, however, when William Stead – author of the book *The Americanization of the World* – took over the editorship of the paper, it moved toward a more American style of journalism (Wiener, 2011; Scott, 1950) [3]. This raises the question to what extent the newspaper re-focused on the United States and whether the notion of the metropolis was re- branded as an American concept.

In this paper, we will show how computational methods such as Named Entity Recognition, Topic Modeling and GIS mapping (see the figure below) can be used alongside traditional close-reading methods to gain insight into ideas of empire and the branding of cities as metropolises. Our hypothesis is that there is a shift from a preoccupation with European metropolises such as Paris, Vienna, and London to cities in the United States such as New York and Chicago.

Firstly, we will employ NER to extract place-names that we will map diachronically over the 30-year period. This will reveal the international outlook of the newspaper and by extension reveals continuities and discontinuities with the public's preoccupation of the English empire in its late mature phase and the beginnings of imperial decline in the late nineteenth century (Go, 2011). Secondly, we will use topic modeling to extract specific events related to location within this period. This information will be compared with the international outlook provided by the NER. Moreover, the output will provide keywords that can point us toward specific articles. Thirdly, we will use a predefined list of European and American cities to determine how they figured in the imagination of the English public. Furthermore, we will employ collocation techniques

as well as specific full-text queries that allow us to study the discursive representation of the city.

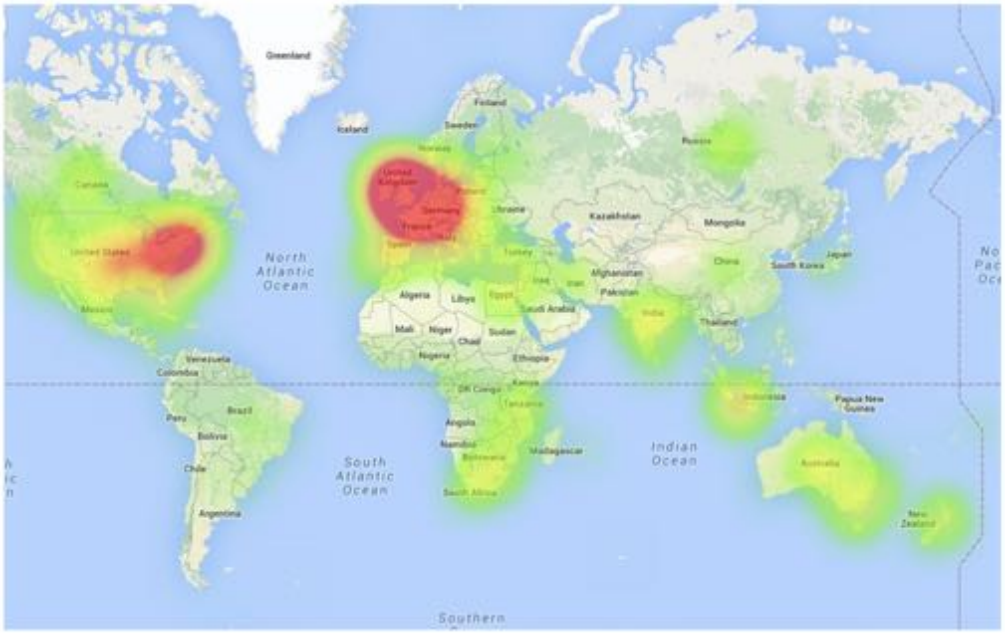


Figure 18: Heat map of locations in the first three months of 1890.

Notes

[1] This period represents a pilot study that can be extended to cover larger periods.
[2] This digitized newspaper is held as part of the British Library newspaper collection.
[3] Historian Joel H. Wiener pointed out that the phrase "Americanization of the British Press" first appeared in the Pall Mall Gazette in 1882.

References

Go, J., 2011. *Patterns of empire: the British and American empires, 1688 to the present*. Cambridge University Press, New York.
Guldi, J., Armitage, D., 2014. *The History Manifesto*. Cambridge University Press, Cambridge.
Hall, C., Rose, S.O., 2006. *At home with the empire: metropolitan culture and the imperial world*. Cambridge University Press, Cambridge.
Körner, A., Smith, A., 2012. *America imagined: images of the United States in nineteenth-century Europe and Latin America*. Palgrave Macmillan, Basingstoke.
Kroes, R., Rydell, R.W., 2005. *Buffalo Bill in Bologna: The Americanization of the world, 1869-1922*. University of Chicago Press, Chicago.
Maier, C.S., 2007. *Among empires American ascendancy and its predecessors*. Harvard University Press, Cambridge, Mass.; London.
Robertson Scott, J.W., 1950. *The story of the Pall Mall gazette, of its first editor Frederick Greenwood and of its founder George Murray Smith*. Oxford University Press, London.
Wiener, J.H., 2011. *The Americanization of the British press, 1830s-1914: speed in the age of transatlantic journalism*. Palgrave Macmillan, New York.

Semantic Enrichment of a Multilingual Archive with Linked Open Data

Max de Wilde
Université libre de Bruxelles
madewild@ulb.ac.be

Simon Hengchen
Université libre de Bruxelles
shengche@ulb.ac.be

The Historische Kranten [1] project involved the digitisation, OCR and online publication of over a million articles from 41 Belgian newspapers published between 1818 and 1972. Articles are written in Dutch, French and English and focus mainly on the city of Ypres and its neighbourhood. Currently, only full-text indexing has been performed on the collection, which means that search for particular mentions in the corpus suffer from both noise and silence. For instance, a search on the string "Huygens" returns correct results about Christiaan Huygens:

> *Links zien wij Christiaan Huygens die met zijn slingeruurwerk de oplossing bracht voor het meten van de tijd*

But one also gets results that are not relevant in this context (noise):

> *La reconnaissance du cadavre de la veuve Huygens, faite par les hommes de l'art, a fait constater l'existence de neuf blessures sur la tête*

Moreover, interesting results are lost due to variations in spelling (silence):

> *en op het uurwerk toegepast door den Hollander Huyghens (1629-1695).*

We first performed Named-Entity Recognition (NER) on this collection in order to extract meaningful concepts. A second step involved a new approach to Entity Linking with gazetteers (Shen et al., 2014) in order to disambiguate them with DBpedia URIs2 (Bizer et al., 2009). For instance, `http://dbpedia.org/resource/Christiaan_Huygens` includes the alternative label "Christian Huyghens" (French spelling) but excludes information about the Belgian painter Léon Huygens (which has his own unique URI: `http://dbpedia.org/resource/Léon_Huygens`) or the crater on Mars named after the Dutch astronomer (`http://dbpedia.org/resource/Huygens_(crater)`).

We now intend to integrate our findings into the project's web interface in order to improve the search experience of the end-users. We plan to interact with the users to get feedback about the relevance of entities extracted and of automatic related search suggestions based on semantic relatedness, which are currently quite random and of poor quality. The impact of OCR quality on NER output will also be evaluated. In a similar experiment on Holocaust-related archives, Rodriquez et al. (2012) find, somewhat counter-intuitively, that "manual correction of OCR output does not significantly improve the performance of named-entity extraction". The confirmation of this hypothesis would mean a lot to institutions that lack sufficient funding to perform first-rate OCR on their collections.

Notes

[1] `http://www.historischekranten.be/`.
[2] We use DBpedia as an entry point to the Linked Data cloud, enabling access to other resources with the `owl:sameAs` property.

References

Bizer, C., Lehman, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – A crystallization point for the Web of Data. *Web Semantics: science, services and agents on the World Wide Web*, 7(3):154–165.
Rodriquez, K. J., Bryant, M., Blanke, T., and Luszczynska, M. (2012). Comparison of Named Entity Recognition tools for raw OCR text. In *Proceedings of KONVENS 2012*, pp 410–414. Vienna.
Shen, W., Wang, J., and Han, J. (2014). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*.

Specifying a system for collaborative online scholarly editing

Peter Robinson
University of Saskatchewan
peter.robinson@usask.ca

There has been considerable interest and activity over the last few years in the making of online collaborative editing systems (Causer, Tonra and Wallace 2012; Rito Silva and Portela 2014). Through the internet, we can open up many scholarly editing activities (particularly, indexing and transcribing manuscripts) to anyone with a computer. This promises that editorial tasks which long seemed impossible (the transcription of all 5000 manuscripts of the New Testament, or all 800 of Dante's Commedia) might now be addressed. Accordingly, many groups have set out to create collaborative online systems, prompted by the landmark success of Transcribe Bentham and similar enterprises: a survey by Ben Brumfield in 2012 listed some 35 tools, and there are many more developed since. The success of the new Text Encoding Initiative Guidelines on Representation of Primary Sources has further encouraged numerous new initiatives in transcription of authorial manuscripts, such as the Shelley-Godwin archive.

However, much of this ferment of activity is based on a limited and flawed model of texts, documents and works. All the systems described by Brumfield, and many others which have arisen since, are "document oriented", even "page oriented". They are superb at representing the text line by line, a page at a time, within a single document. However, they are unable to record the structure of the work: its division into hierarchies of section, chapter and paragraph, or stanzas and verses. As a result, their use in editions which seek to record differences and agreements between instances of the work in separate documents, or even differences which span across page barriers, is highly problematic (Muñoz, Viglianti and Fraistat 2013; Bruning, Henzel and Pravida 2013). We call the structured, ordered and hierarchical divisions of a work "entities", corresponding to the familiar TEI hierarchy of (for example) text, body, division, paragraph.

This paper will survey the problems which arise from "page oriented" transcription and will demonstrate a system developed by the presenter and partners, Textual Communities, which seeks to address these limitations. This system is now in use on some six major projects, and currently holds some 40,000 images and transcripts of manuscript pages spread across around 200 documents, with around one hundred and fifty transcribers at work on them. The system is built on a precisely formulated ontology of documents, entities, and texts, which both maps the two aspects of texts as "page" and "entity" onto distinct ordered and structured hierarchies, and then interlinks the hierarchies so that one may readily view (and edit) a text by page or by entity (first figure below). One may, for example, readily extract all the different versions of (say) the first line of the General Prologue of the Canterbury Tales and send them to a collation tool (CollateX; see the second figure below). The heart of Textual Communities is a set of routines for managing these multiple aspects of documents through use of TEI-native technologies. In our most recent iteration of the Textual Community

database backend we are moving the fundamental document materials into a JSON-based storage system, MongoDB. In the process, we have found that we are able both to support an unlimited number of overlapping hierarchies beyond the two (of document and entity) that we currently enable, and also to represent each hierarchy more richly than we are currently able to do. This opens interesting possibilities for moving past the difficulties XML has with coping with multiple hierarchies.

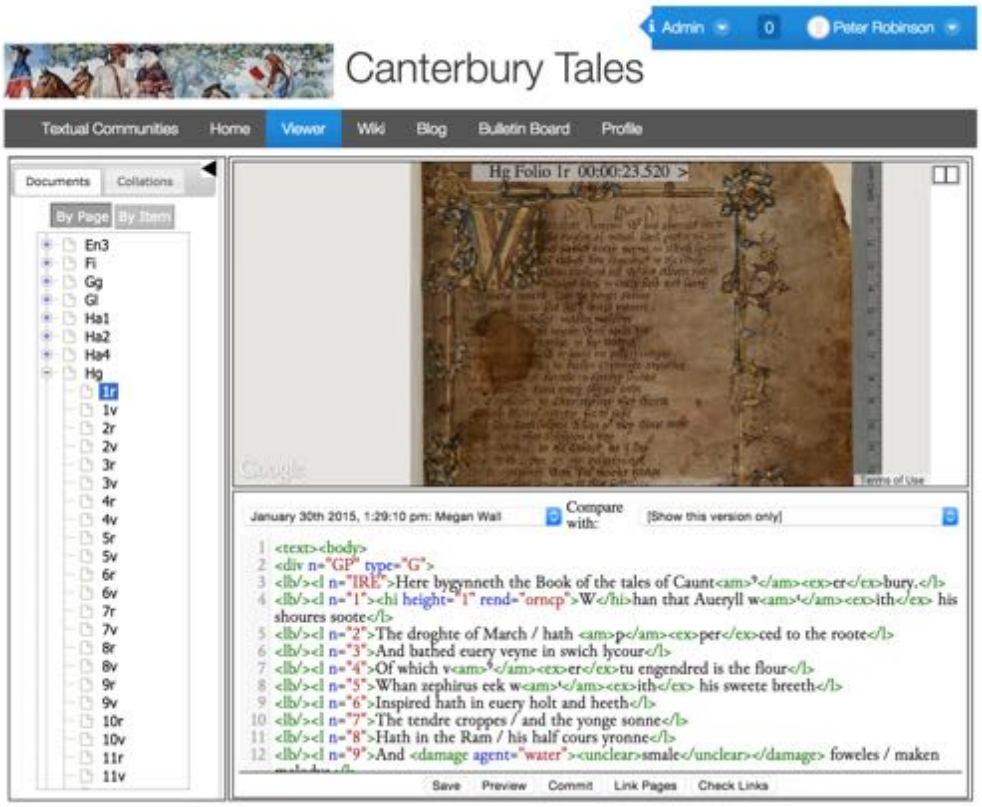


Figure 19: Textual Communities, showing view by document and page, with editing environment.

Textual Communities is itself both available as open-source, and is itself completely built on open-source software, accessible at `https://github.com/DigitalResearchCentre`. The system itself may be currently viewed at `http://www.textualcommunities.usask.ca/`; the Canterbury Tales project materials in that site exemplify well its abilities. Although currently open for anyone to use, we are testing it thoroughly and refining the interface before announcing full public release. It has been developed with support from the University of Saskatchewan, the Canadian Foundation for Innovation, and the Social Science and Humanities Research Council of Canada.

References

Brumfield, B. (2012) 'Crowdsourced Transcription Tool List', April 11 2012, `http://tinyurl.com/TranscriptionToolGDoc`.
Brüning, G, Henzel K., and Pravida, D. (2013.) "Multiple Encoding in Genetic Editions: The Case of 'Faust'." *Journal of the Text Encoding Initiative* [Online] (`http://jtei.revues.org/697`).

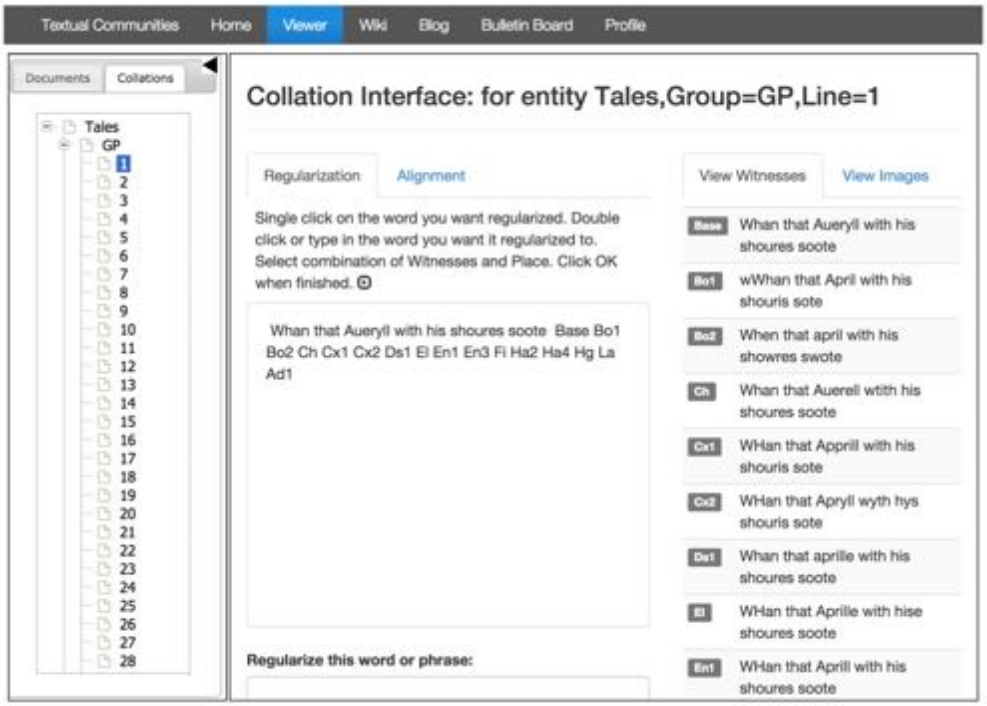


Figure 20: Textual Communities, showing output of CollateX comparison of sixteen versions of the first line of the General Prologue of the Canterbury Tales, after regularization using tools within this interface..

Causer, T., Tonra, J. and Wallace, V. (2012) 'Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham, *Lit. Linguist Computing* 27 (2): 119-137.

Robinson, P. M. W. (2012) 'Towards a Theory of Digital Editions'. *Variants* 10: 105-131 (`https://www.academia.edu/3233227/Towards_a_Theory_of_Digital_Editions`).

Maryland Institute for Technology in the Humanities, et al. *Shelley-Godwin archive* (`http://shelleygodwinarchive.org/`.

Rito Silva, A. and Portela, M. (2104). 'TEI4LdoD: Textual Encoding and Social Editing in Web 2.0 Environments', *Journal of the Text Encoding Initiative* [Online] (`http://jtei.revues.org/1171`.

Text Encoding Initiative. P5. (2007–): Guidelines for Electronic Text Encoding and Interchange. Chapter 11. "Representation of Primary Sources" (`http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html`).

TEI Annotation and Network Analysis of Diplomatic History Documents

Florentina Armaselu
CVCE Luxembourg
florentina.armaselu@cvce.eu

Marten Düring
CVCE Luxembourg
marten.during@cvce.eu

Veronica Martins
CVCE Luxembourg
veronica.martins@cvce.eu

In this paper we explore the potential of network analysis for the exploration of TEI annotated documents. Network visualizations provide insight in highly complex relations between any type of entities such as individuals or locations. Visualizations are used to reveal patterns in data which are otherwise impossible or very hard to detect. We will discuss three use cases: 1) networks as means to create an interlinked taxonomy in which hierarchies are expressed through centrality scores. 2) Centrality algorithms describe structural properties of nodes and networks mathematically. We identify possible synonyms based on the axiom of homophily which is also entailed in the proverb that "birds of a feather flock together". 3) We discuss the added analytical value of network visualizations for a domain expert in diplomatic history with in-depth knowledge of the underlying primary sources.

The studied corpus is part of a larger research project on the diplomacy within Western European Union (W.E.U.) and contains documents (French) on the production, standardisation and control of armaments (1954 to 1982) from the W.E.U. archives which are based at the Archives Nationales de Luxembourg. It includes different types of materials encoded in XML-TEI P5, e.g. notes from the Secretary-General or Secretariat-General, minutes of meetings, memoranda and studies. Three categories of encoding are provided: metadata (title, author, availability date, origin place, confidentiality status, etc.), structural markup (headers, footers, sections, paragraphs, line breaks), content-related annotations (discourse of country/institutional representatives, named entities). The Named Entity Recognition (NER) task involved a semi-automatic approach using GATE, i.e. the French NE system, Gazetteer and Gazetteer List Collector plugins. Seven classes of entities were identified and annotated in the texts: persons, places, organisations, events, dates, products and functions (official positions).

Through network analysis via Gephi, we will address questions related to topics like: use of variants for the same entity, the so-called "synonyms" in the context (Union de l'Europe occidentale / Union de l'Europe Occidentale / UNION DE L'EUROPE OCCIDENTALE / U.E.O. / U. E. O.); hierarchical relations between entities (Conseil / Conseil ministériel / Conseil ministériel de l'U.E.O.); interpretation from a historical perspective of the different types of networks built with the annotated data (persons, organisations, etc.).

References

GATE (General Architecture for Text Engineering) [WWW Document], https://gate.ac.uk/ (accessed 3.30.15).
Gephi (The Open Graph Viz Platform) [WWW Document], http://gephi.github.io/ (accessed 3.30.15).
TEI (Text Encoding Initiative) [WWW Document], http://www.tei-c.org/index.xml (accessed 3.30.15).

The Digital Humanities cycle: hermeneutics, heuristics, and source criticism in a digital age

Jesper Verhoef
Utrecht University
j.verhoef@uu.nl

Melvin Wevers
Utrecht University
m.j.h.f.wevers@uu.nl

This paper demonstrates how we have applied a question-driven approach to digital history using computational methods. We use the analogy of the Digital Humanities Cycle – an updated version of the empirical cycle – to explain the iterative process of heuristics, hermeneutics, tool criticism, corpus faceting, and source criticism. The Digital Humanities Cycle entails that, while doing historical research, the strength of numerous digital tools and archives should constantly be combined and alternated with 'classical' historical hermeneutics, i.e. source criticism and interpretation. The Figure below shows a visualization of the way in which the Cycle functions.

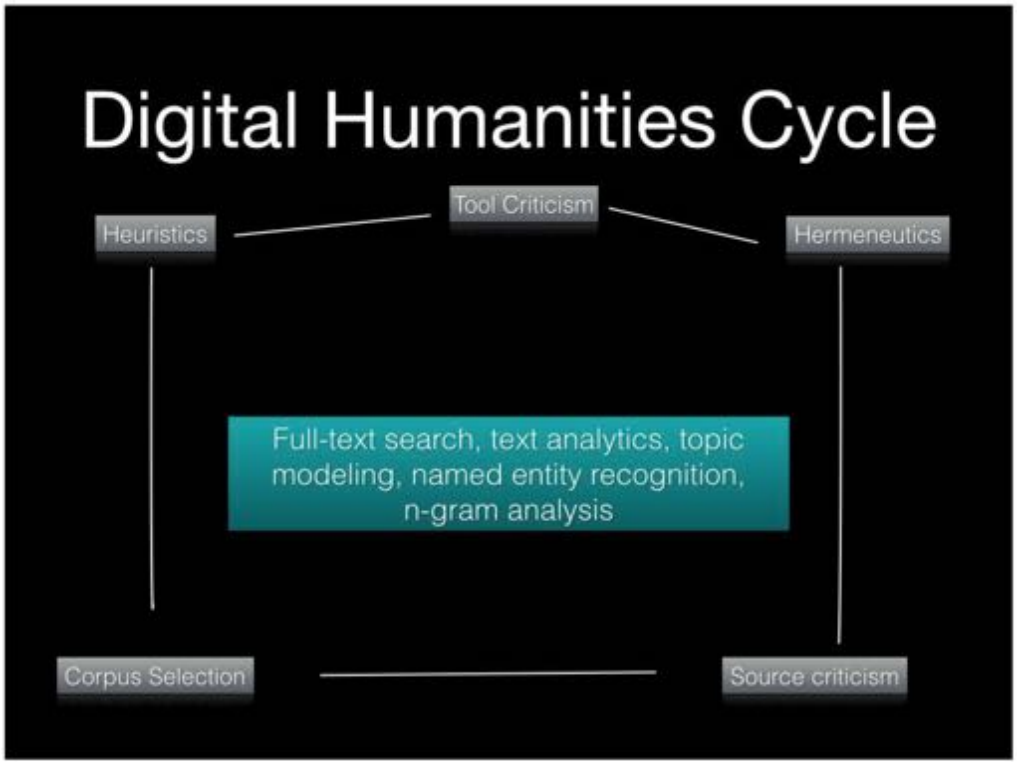


Figure 21: The Digital Humanities Cycle

.

We will briefly touch upon each of these aspects through examples from our own research that deal with the representation of America in Dutch digitized newspa-

pers. These examples include debates on transistor radios and the construction of geographic connotations of consumer goods.

In our research, we have applied the Cycle to the digitized repository of the National Library of the Netherlands, which contains over 600,000 digitized newspapers between 1890 and 1990. The paper describes how full-text search, N-gram analysis, Topic Modeling, and Named Entity Recognition help to select (sub-)corpora of sources, and generate themes and topics of interest to the study. More importantly, we will show that the stages in the Cycle – searching the archive (heuristics), understanding whether and how a tool works, confining a corpus of articles and applying corpus linguistics and source criticism, and interpreting the results (hermeneutics) – are all interrelated. For instance, when applying topic modeling on all the newspapers articles that contain the word 'portable radio' – a result of full text search in the first place –, the topics which are produced demand interpretation. Moreover, they render new (full text) search words, and at the same time – when topics clearly demarcate a single theme – may lead a researcher to apply other digital tools to this very dataset. This could include corpus linguistic tools to clean up the corpus or extract other meaningful linguistic information. Put differently, the output of one single tool is never the endpoint of research, but one of the many step needed to weave a historical narrative.

The Digital Humanities Cycle is a response to the plea made by numerous scholars that, after the Digital or Computation Turn (Nicholson 2013; Berry 2011), one should obtain 'hybridity', that is combine classic and digital historical research (Zaagsma 2013). When applied in combination, these techniques might yield patterns in the dataset, without missing "the power of the particular" (Hitchcock 2014). We will show that the Digital Humanities Cycle leads to a fruitful cooperation between "counting and understanding" (Rieder & Röhle 2012) – one of the prominent challenges when using digital tools.

References

Berry D.M. (2014), 'The computational turn: thinking about the digital humanities', *Culture Machine*, vol. 12, pp. 1–22.

Hitchcock T. (2014), 'Big data, small data and meaning' on *Historyonics* (Blog). Available from: `http://historyonics.blogspot.co.uk/2014/11/big-data-small-data-and-meaning_9.html` [13 November 2014].

Nicholson B. (2013), 'The digital turn', *Media History*, vol. 19, no. 1, pp. 59–73.

Rieder, B. and Röhle T. (2012), 'Digital methods: five challenges' in *Understanding Digital Humanities*, ed. D.M. Berry, Houndmills: Palgrave Macmillan, pp. 67–84.

Zaagsma Z. (2013), 'On digital history', *BMGN - Low Countries Historical Review* vol. 128, no. 4. pp.3–29.

The Document as Event. Applying automatic semantic role labeling to collections of theatre reviews

Thomas Crombez
Royal Academy of Fine Arts Antwerp
St. Lucas University College of Arts (Antwerp)
University of Antwerp
thomas.crombez@uantwerpen.be

When a theatrical performance is over, nothing remains but the memories of the participants and witnesses. At least, that is the well established cliché in the performing arts, and in theatre history as well. To investigate a theatrical event is, by definition, to investigate a historical event. It implies for that matter the study of the event's traces: the material record of costumes, props and set designs; the written scenario or script (when available); and the eyewitness accounts of performers, technicians and spectators.

For theatre historians, there is an important extra source: the critical record, consisting of reviews and essays. Historical documents of theatre criticism can bring a former performance context and its related practices (acting styles, physical communication, audience reactions) back to life. Only recently have such documents become accessible through large-scale digital collections.

This paper explores the possibilities of extracting information about historical performances from large collections of theatre reviews. In the literature, this is known as a subgenre of text-mining, namely, (semi-)automatic semantic role labeling (Gildea and Jurafsky 2002). The resulting data are an invaluable source of information for theatre history. Not only do they constitute a virtual record of past theatrical events, but they may also be used for studying processes of canonization (e.g., which names rise and fall in frequency?) and changing tendencies in the performing arts (e.g., the diminishing importance of the dramatist in the second half of the 20th century).

Two techniques from natural language processing and machine learning are crucial in the process of transforming plain-text documents into semantically marked-up data. First, all named entities should be detected and extracted from the text. Second, the extracted names have to be classified according to their specific role. In my paper, I will focus on detecting artistic roles of personal names in the context of the performing arts (i.e., roles such as dramatist, director, performer, scenographer, or critic), based on the surrounding snippets of text. Further, I will demonstrate the results and possible use-cases based on mining documents from different collections of theatre reviews, coming from Belgian newspapers in French and Dutch from the 1989–2014 period.

References

Gildea, D. and Jurafsky, D. (2002) 'Automatic Labeling of Semantic Roles', *Computational Linguistics*, vol. 28, 3, pp. 1–45.

The History of Cultural Heritage Collections: Exploring New Approaches to Analysis and Visualization

Toby Burrows
King's College London
toby.burrows@kcl.ac.uk

In the nineteenth century, the English collector Sir Thomas Phillipps (1792-1872) assembled the largest private collection of European medieval and early modern manuscripts and documents. It is estimated to have contained more than 40,000 items, making it considerably larger than most of the collections in public institutions today, and included many manuscripts of considerable historical, textual and artistic significance. Their modern locations are spread across the globe – the dispersal of the Phillipps Collection took place gradually over more than one hundred years, and numerous institutions and collectors were involved. As a result, the Phillipps Collection provides a rich and varied set of data for tracking the history and provenance of cultural heritage collections.

In this paper, I will report on a project to reconstruct and analyse the history and provenance of the manuscripts which formed the Phillipps Collection. The scale of the Phillipps Collection has proved a significant challenge to traditional research methods in the past; the English librarian A.N.L. Munby spent more than a decade compiling a overview of Phillipps' collecting activities and of the dispersal of the collection up to the mid-1950s (Munby 1951-1960). In this project I am employing data modeling and analysis techniques to build a digital environment for tracing the entire history of these manuscripts, as far as it can be known. I am interested in mapping the provenance events and ownership networks which, taken together, constitute the history of these thousands of manuscripts over hundreds of years.

My paper will focus on four key technical aspects of the project.

1. *Frameworks for modeling and representing the data relating to ownership and provenance, using an event-based approach*. Events are central to provenance research, but they have proved difficult to represent in existing ontologies and data models, with a variety of different approaches being used. I will discuss some of these – including CIDOC-CRM, the Europeana Data Model, and property graphs.

2. *Techniques for importing and combining existing data relating to manuscript histories*. The existing data relating to the Phillipps manuscripts range from relational databases and MARC records to handwritten notes and card indexes. Capturing and cleaning these data and aligning them to a common data model are complex tasks which require multiple ingestion paths and crosswalks.

3. *The deployment of suitable software to manage the data and to support analysis and visualization*. I will report on two specific platforms: the graph database software Neo4j (Van Bruggen 2014) and the nodegoat data management environment (Van Bree and Kessels 2015).

4. *Methods for visualizing and analyzing the data produced by the project, and for making them available for re-use by other researchers*.

I will look at a series of use cases and research questions related to the aggregated data, and will demonstrate how Neo4j and Nodegoat can be use to produce analyses and visualizations in response to these requirements. I will also discuss methods for linking the data produced by this project with the wider Linked Data cloud, in order to enable wider contextualization and analysis.

References

Munby, A.N.L. (1951-60) *Phillipps Studies*, Cambridge: Cambridge University Press. 5 vols.

Van Bree, P. and Kessels, G. (2015) "Mapping memory landscapes in nodegoat" in: *Social Informatics*, ed. L.M. Aiello and D. McFarland (Lecture Notes in Computer Science 8852), pp. 274-278, New York: Springer International.

Van Bruggen, R. (2014) *Learning Neo4j*, Birmingham: Packt Publishing.

The TRAME Project – Text and Manuscript Transmission of the Middle Ages in Europe

Emiliano Degl'Innocenti
Fondazione Ezio Franceschini
emiliano@fefonlus.it

Alfredo Cosco
SISMEL, Firenze
alfredo.cosco@gmail.com

TRAME, Text and Manuscript Transmission of the Middle Ages in Europe, is a research infrastructure for medieval manuscripts, hosted by FEF/SISMEL on: http://git-trame.fefonlus.it. The project was born in 2011. Its main aim was to build a research infrastructure project focused on promoting interoperability among different digital resources available in the medieval digital ecosystem by connecting repositories of digitized images of medieval manuscripts, their codicological descriptions, their textual and philological interest and their cultural significance in the context of the European history.

Currently it implements a number of features (including simple, shelfmark or advanced search mode etc.) on more than 80 selected scholarly digital resources on western medieval manuscripts, authors, and texts across the EU and USA, including digital libraries, research databases and other projects from leading institutions. TRAME is more than just a piece of software: it is a research tool deeply rooted in the international medieval scholarly community, whose development is in line with the Memorandum of Understanding of the COST Action IS1005 "Medieval Europe Medieval Cultures and Technological Resources", representing 260 researchers coming from 39 leading institutions (archives, libraries, universities and research centers) in 24 countries across EU. The engine scans a set of sources for searched query terms and retrieves links to a wide range of possible information, from simple reference to detailed manuscript record, to full text transcriptions.

The application is written in OO-PHP, the design follows the HMVC Pattern, the RDBMS is MySql and the front-end combines Xhtml and Javascript. Since 2014 September 1st, TRAME has entered a new phase and the current work is focused on:

- extending the meta-search approach to other web resources,
- leveraging the users interaction to define an ontology for medieval manuscripts,
- re-designing the front-end towards a new UX approach.

TRAME is an ongoing collaborative international effort, rooted in the medieval research community. Its development agenda is deeply influenced by the needs expressed by scholars across EU and US. Recent changes about the nature of the information available in the WWW influenced the development of TRAME from a mere meta search approach towards the establishment of a Medieval Semantic Knowledge base, using custom modules for information collection and integration (i.e. web crawler, data miner).

References

E. Degl'Innocenti, Trame: Building a Meta Search Tool for the Study of Medieval Literary Traditions in *EVA 2011, Proceedings*. Vito Cappellini, James Hemsley (eds.), Bologna, Pitagora, 2011.

TRAME. Text and Manuscript Transmission of the Middle Ages in Europe. Evolving the System Towards Horizon2020 and VCMS Challenges. http://www.sismelfirenze.it/index.php?%option=com_k2&view=item&task=download&id=68_69e648d4f36e436d0ec96c334a01%80a4&Itemid=266&lang=it.

The curation of sound archives: the Dutch Dialect Database

Douwe Zeldenrust
Meertens Institute KNAW
douwe.zeldenrust@meertens.knaw.nl

Introduction

At the first DH Benelux conference, papers and presentations concerning the digitization, curation and dissemination of digital objects were in a minority. Within this domain the interests in sound archives and the use of sound recordings as a resource were even more limited and restricted to for instance phonologists and musicologists. But in a recent article in *Digital Humanities Quarterly*, Anne Murray and Jared Wiercinksy placed this subject in a broader perspective. They state that the creation of new knowledge in the humanities 'depends not only on better understanding the role of sound in the work of humanities scholars, but also incorporating this knowledge into the design of sound archives' (Murray).

Use case: the Dutch Dialect Database

This paper focuses on the digitization, curation and design of sound archives. A use case will be presented: the project of the Dutch Dialect Database (NDB) of the Meertens Institute, Royal Netherlands Academy of Arts and Sciences (Oostendorp). The project started in 2009 and was presented at the 2014 DHBenelux conference as the 'Soundbites collection' (Zeldenrust). Since then, the digital sound archive has been renewed significantly. The latest design principles for web-based spoken word archives have been incorporated and extra information has been added. Currently the collection contains Dutch spoken in the Netherlands and also Dutch spoken outside the Netherlands, such as Dutch spoken in the USA and Dutch spoken in France (Archives). The datasets are CLARIN (Common Language Resources and Technology Infrastructure) compatible and are available via the CLARIN interface.

Conclusion

To conclude with, the paper will reflect on the creation, curation and dissemination of humanities digital resources. The combination of sound and written resources with specially designed interfaces and tools will bring new research possibilities. For instance, the availability the collection Dutch Spoken in the USA will open the way for pattern recognition within this dataset (Sijs).

References

Murray and Jared Wiercinski (2014) 'A Design Methodology for Web-based Sound Archives', *Digital Humanities Quarterly*, Volume 8 Number 2.

Oostendorp, M. van (2014), 'Phonological and phonetic databases at the Meertens Institute', *The Oxford Handbook of Corpus Phonology*. Eds. J. Durand & G. Kristoffersen. Oxford: OUP. 546–551.

Sijs, N. van der. (2013) 'Op de schouders van onze voorgangers. Jo Daan en het Amerikaans-Nederlands', *Nieuw Letterkundig Magazijn* 31.1: 30–34.

Zeldenrust, D.A. (2014) 'Access to data: the Soundbites collection of the Meertens Institute and a flexible approach to the curation and dissemination of humanities digital resources?, *Digital Humanities Benelux Conference* (The Hague).

Archives
The Archives of the Meertens Institute, Nederlands in Frankrijk, (1964-1966).
The Archives of the Meertens Institute, Nederlands in de USA, collection Jo Daan, (1966).
The Archives of the Meertens Institute, Nederlands in de USA, collection van Marle, (1994-1998).

Websites
www.clarin.eu (Accessed February 5, 2015).
www.meertens.knaw.nl/ndb (Accessed February 5, 2015).

The possibilities and challenges of using linked data for academic research: the case of the Talk of Europe project

Laura Hollink
VU University Amsterdam
l.hollink@cwi.nl

Martijn Kleppe
Erasmus University Rotterdam
kleppe@eshhc.eur.nl

Max Kemman
University of Luxembourg
max.kemman@uni.lu

Astrid van Aggelen
VU University Amsterdam
a.e.van.aggelen@vu.nl

Willem Robert van Hage
SynerScope
willem.van.hage@synerscope.com

The Talk of Europe project has made the proceedings of the plenary meetings of the European Parliament available as Linked Open Data, a way of publishing and connecting data on the Web. Access to the records of what happens during the meetings of the European Parliament (EP) is a crucial part of democracy. In addition, the proceedings are valuable source material for scholars in history, politicology (Proksch, 2010), natural language processing (Nusselder, 2009) and machine translation (Koehn, 2005). However, the EP web portal only offers limited search functionality. By publishing this data as Linked Open Data, we aim to improve access for scholars. In this presentation we will reflect on the benefits and implications of linking the proceedings of the EP to a number of other datasets. Additionally, during the demonstration session, we will show how the SynerScope visualization tool enables an exploration of the links within and across datasets.

Up until now, we have linked the proceedings to four external datasets: 1) a database of professional affiliations of the members of the EP (Høyland, 2009), 2) DBpedia [1], the semantic web mirror of Wikipedia, 3) Geonames [2], a geographical thesaurus, and 4) the politicians and parties of the parliament of Italy [3]. Through these links, we enable scholars to access and use the knowledge that is captured in these external datasets. For example: the former member of parliament Jeanine Hennis is linked to her entry in Høyland's database telling us that she was a member of the Committee on Transport and Tourism; to her DBpedia page, giving access to her birthdate and place, diplomas, and jobs outside the EP; the country that she represents (The Netherlands) is linked to its corresponding Geonames entity, providing information on population density, income and neighbouring countries. These links enable queries that are not possible on either of these datasets alone. For example, if we combine birthplace information from DBpedia with Geonames' geographical information, we can query for members of the EP that were born outside Europe.

During two creative camps, we invited teams of scholars to use our linked dataset. Interesting applications were built on top of it, showing the possibilities of Linked Open Data for scholars. However, at the same time, an increase in the number of exter-

nal datasets that is used raises questions about the correctness, transparency, stability and completeness of the results, which are fundamental questions for humanities researchers, for whom provenance is crucial. We furthermore observed that while the organisation of the creative camps stimulated the uptake of the dataset, the Linked Open Data format remains a hurdle for many humanities scholars.

Notes

[1] `http://dbpedia.org/`
[2] `http://www.geonames.org/`
[3] `http://data.camera.it/`

References

Proksch, S.-O. and Slapin, J.B. (2010) Position taking in european parliament speeches. *British Journal of Political Science*, Vol. 40, Issue 03, pp. 587–611.

Nusselder, A., Peetz, H., Schuth, A. and Marx, M. (2009) Helping people to choose for whom to vote. A web information system for the 2009 european elections. *Proceedings of the 18th ACM conference on Information and knowledge management*, ACM, pages 2095–2096.

Koehn, P. (2005) *Europarl: A parallel corpus for statistical machine translation*. MT summit, volume 5, pages 79–86.

Høyland, B., Sircar, I. and Hix, S. (2009) An Automated Database of the European Parliament. *European Union Politics*, Vol 10, Issue 1, pp. 143-152 .

TheRiddlerBot: A Next Step on the Ladder Towards Computational Creativity

Ben Verhoeven
University of Antwerp
ben.verhoeven@uantwerpen.be

Iván Guerrero
Universidad Nacional Autónoma de México
ivangro@gmail.com

Francesco Barbieri
TALN, Universitat Pompeu Fabra
francesco.barbieri@upf.edu

Pedro Martins
Universidade de Coimbra
pjmm@dei.uc.pt

Rafael Pérez y Pérez
Universidad Autónoma Metropolitana, Cuajimalpa
rperez@correo.cua.uam.mx

We present a computational system that aspires to be considered creative by generating riddles about celebrities and well-known characters. The riddles are crafted by combining information from both well-structured and poorly structured information sources. This system has been implemented as an interactive twitter bot (@TheRiddlerBot). We intend our bot to go beyond mere generation of tweets where it makes all possible combinations of the information in our data, rather we want the riddle to be correct, coherent, and fun to solve. Each iteration of the riddle creation process involves selecting a character/person from the database – preferably not random, this could be based on current events extracted from news websites – and gathering knowledge about him/her by combining both our knowledge base and online resources, e.g. Wikipedia. The challenge then lies in selecting the right and appropriate information and presenting it in a fun and hopefully non-obvious way. One way of doing this is creating analogies between users by means of common attributes, e.g. 'Doc Emmett Brown' could be 'the Walter White' of 'Back to the Future' given that they are both scientists. The information is formed into a riddle by making extensive use of a multitude of phrase templates. In the future we would like to be able to validate each aspect of the riddle creation in order to derandomize all modules and thus having all choices made 'consciously' by the system. For that purpose – and for the fun of it – we have made our twitter bot interactive. People can guess whom the riddle is about and reply accordingly. The first right answer for each riddle is rewarded with a point. If your answer is incorrect, the bot will tell you so. In the mean time, our system is gathering data about the number of answers, favorites and retweets in the hope that these might tell us something about the quality of the riddles.

Thomas Kling's Bakchische Epiphanien – Projekt "Vorzeitbelebung" Reconstructing the Digital Writing Process from a Hard Drive in the Thomas Kling Archive (Digital Forensics)

Thorsten Ries
Ghent University
thorsten.ries@ugent.be

The talk will outline the philological reconstruction of the digital writing process of Thomas Kling's *Bakchische Epiphanien* essays, titled *Projekt "Vorzeitbelebung"* with digital forensic tools and methods, and discuss examples from the digital dossier gén'etique from a textual genetic as well as scholarly editing perspective. The born digital draft materials, temporary files, textual data fragments and other writing process traces of *Projekt "Vorzeitbelebung"* have been recovered from a forensic image of a hard drive in the collections of the Thomas Kling Archive (Stiftung Insel Hombroich) in the course of an FWO funded research project. In the poetological essay, published 2005 in Kling's last volume of poetry and essays *Auswertung der Flugdaten*, Kling probes into poetic modes of intertextuality and poetological reflective representation of the poetic cultural archive (*"Poetik, Archivbilder"*, *"Bakchische Epiphanien"*) throughout his personal canon which in this case ranges from Euripides, Ovidius, the ancient cults of Dionysus Bromius, Hermes and shaman traditions to Stefan George, Rudolf Borchardt, Ortega y Gasset, Ezra Pound and Gottfried Benn. Tracing the digital writing process of this short essay collection with digital forensic tools not only demonstrates digital forensic methods relevant for all archival studies on born digital material, but also sheds light on the intellectual composition of Kling's reasoning on poetological intertextuality (*"Sondage"*).

The talk will, on the methodological level, discuss the application of forensic computer science tools and methods to born digital documents and parts of archives, focusing on the philological benefit for genetic scholarly editions and the critique gén'etique on the one hand as well as on issues of sane archiving and representation of the digital record in a scholarly edition on the other. The talk will try and map the conceptual impact of the digital forensic perspective and the concrete materiality of the digital historical record on philologically central terms and concepts such as 'document', 'materiality', 'writing process', 'trace' as well as on the practice of (documentary) scholarly editing.

References

Ammon, Frieder von: *Von Epenchefs und Studienabbrechern. Zur Essayistik Thomas Klings*. In: Frieder von Ammon, Peer Trilcke, Alena Scharfschwert (Eds.) *Das Gellen der Tinte. Zum Werk Thomas Klings*. Göttingen: V&R unipress 2012, 41-67.

Burdorf, Dieter. *Bakchenrasereien. Thomas Kling liest Rudolf Borchardt*. Uta Degner, Elisabetta Mengaldo (Ed.) *Der Dichter und sein Schatten. Emphatische Intertextualität in der modernen Lyrik*. München: Fink 2014, p. 117-134.

Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine (Eds.): *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Council on Library and Information Resources Washington, D.C. December 2010 (CLIR publication, no. 149).

Kirschenbaum, Matthew G. 'The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary'. In: *Digital Humanities Quarterly* 2013.7.1 (http://www.digitalhumanities.org/dhq/vol/7/1/000151/000151.html) accessed: February 28, 2015.
Kling, Thomas: *Rhapsoden am Sepik*. Botenstoffe. Cologne 2001.
Kling, Thomas: *Auswertung der Flugdaten*. Köln: Dumont 2005.
Ries, Thorsten."die geräte klüger als ihre besitzer" Philologische Durchblicke hinter die Schreibszene des Graphical User Interface. Überlegungen zur digitalen Quellenphilologie, mit einer textgenetischen Studie zu Michael Speiers "ausfahrt st. nazaire". *editio* 24.1 (2010), 149-199.
Trilcke, Peer: *Historisches Rauschen. Das geschichtslyrische Werk Thomas Klings. [The Historical Poetry of Thomas Kling]*. PhD thesis Göttingen University 2011/12 (http://hdl.handle.net/11858/00-1735-0000-0006-AEDE-3), accessed: October 30, 2013.

Topics in Dutch Minority Languages on Twitter

Anna Katrine Jørgensen
Meertens Institute
anna.jorgensen@meertens.knaw.nl

Lysbeth Jongbloed-Faber
Fryske Akademy
ljongbloed@fryske-akademy.nl

Jolie van Loo
Meertens Institute
j.j.e.vanloo@gmail.com

In this paper, we focus on the use of two Dutch minority languages, Frisian and Limburgian, on Twitter and the effect of topics on choosing language. Almost nine out of ten Dutch people are active on social media, with Twitter claiming 3.5 million tweeters from the Netherlands [2]. Furthermore, over 10% of tweeters tweet in more than one language [3]. Frisian is the mother tongue of approximately half of the Frisian population. The majority of the 650,000 inhabitants of Fryslân can understand the language (very) well (85%), 64% of the population can speak it (very) well, while only 12% indicate that they can write it (very) well [4]. Limburgian mainly functions as a spoken language and is highly variational. 70-75% of the population in Limburg can speak the local dialect and speakers from all socio-ecomonic levels do so in a variety of settings [5]. Interestingly, both languages are used on Twitter, despite their mainly spoken quality. A Frisian or Limburgian bi- or multilingual's choice of language and variety from her linguistic repertoire relies on a number of factors, including the audience [6], offline language use with peers [7], attitude [7], writing skills [7], the perspective, the tone of the tweet and the topic [1,6].

This paper presents an interdisciplinary study of the use of topics by Frisian and Limburgian Twitter users. We present a large-scale, rule-based classification of topics in these minority languages, and discourse about the promises and pitfalls of automatised topic classification on Twitter in this setting. Furthermore we describe the influence of topic on the (de-)selection of a language variety from a tweeters repertoire. The reasons for choosing a minority language for a certain topic are debated and reasons hereof provided. We discuss the differences between the Frisian and the Limburgian distributions of topics, both in their use of Dutch and in the differences between the distributions in the two minority languages. This interdisciplinary approach of data-driven topic modelling and sociolinguistics provides a detailed overview of the impact of topic on the use of minority languages on Twitter. The strength of the approach is two-fold: the sociolinguistic study is generalisable and relies on the analysis of large data sets, while the output of the rule-based topic classifier is subject to thorough investigation and is applied insights from sociolinguistic theory.

This paper is the result of the interdisciplinary project 'Twidentity' which involves Meertens Instituut, Universiteit van Twente, University Maastricht and Fryske Akademy. The Twidentity team consists of Anna Katrine Jørgensen, Jolie van Loo, Lysbeth Jongbloed-Faber, Leonie Cornips, Theo Meder, Dong Nguyen and Dolf Trieschnigg.

References

[1] Androutsopoulos, J. *Code-switching in computer-mediated communication. Pragmatics of Computer-mediated Communication*. De Gruyter Mouton. 2013.
[2] Boekee, Steven et al., *National Social Media Survey 2014*. Newcom Research Consultancy B.V., Amsterdam 2014.
[3] Hale, S.A., Global connectivity and multilinguals in the Twitternetwork. Proceedings of *CHI 2014*.
[4] Province of Fryslân. *Friese Taalatlas*: `http://www.fryslan.frl/taalatlas,2011`).
[5] The Limburgish Academy Foundation. *Limbugs, Modern Usage*, 2009: `http://www.limburgs.org/en/limburgish/modern-usage`.
[6] Nguyen, Dong et al. Audience and the Use of Minority Languages on Twitter. Forthcoming, 2015.
[7] Jongbloed-Faber, Lysbeth et al., The Impact of Social Media on Language Vitality: Online Practices of Bilingual Teenagers in Fryslân. Forthcoming, 2015.

Tourist or Pilgrim? Modeling two types of travel bloggers

Tom van Nuenen
Tilburg University
tomvannuenen@gmail.com

Suzanne van der Beek
Tilburg University
S.E.vdrBeek@uvt.nl

The typological distinction between the pilgrim and the tourist has often been drawn in tourism studies, either theoretically or through ethnographic research (Cohen 1979, Knox et al. 2014). The current paper aims at complementing said debate by adopting a macro-perspective, applying computational stylistic techniques to analyze discursive differences in a corpus of about 7000 blogs from the Dutch travel blog repository of `waarbenjij.nu`. The hypothesis is that tourists and pilgrims share notable similarities in their narratives.

Two sub-corpora are scraped and considered, pertaining to respectively narratives by pilgrims traveling to Santiago de Compostela and by tourists visiting New York City. The great diversity of New York tourists mirrors the diversity in pilgrims found on the Camino, who can travel to Santiago with a variety of backgrounds, expectations, modes of transportation, and amount of time to spend. This term was chosen to ensure that the corpora would consist of texts about journeys that are structurally dissimilar, to capture the important difference in the conception of one's destination between a pilgrim and a tourist.

Several unsupervised computational representations (most notably document-term matrices and topic models; Jockers 2013) are leveraged to analyze the corpora, which yields a bottom-up, data-driven perspective on the differences between these traveler types. A model of ten topics is created based on the part-of-speech tagged corpora, including only nouns. These topics provide a cue to recontextualize the differences found through quantitative analysis. Further interpretation will rely on the application of a qualitative close reading into the indicated themes (Ramsay 2011). The analysis shows that pilgrims, in contrast to theory, write far less about sacred topics and much more about the same practical topics that tourists are stereotypically invested in. Conversely, tourists show a notable sensibility to a wide range of highly valued, set apart experiences. The paper ends with the proposal for a new continuum to understand the difference between tourists and pilgrims, based on the continuum between 'condensed diversity' and 'extended engagement'. By doing so, it aims to contribute to both the understanding of the much debated typology of the two figures, as well as to the field of methodological strategies within the humanities.

References

Cohen, E., 1979. 'A Phenomenology of Tourist Experiences'. *Sociology*, 13:2, pp.179-201.

Jockers, M.L., 2013. *Macroanalysis: Digital methods and literary history*, Champaign, IL: University of Illinois Press.

D. Knox & K. Hannam: 'The secular pilgrim: are we flogging a dead metaphor?', in *Tourism Recreation Research* 39:2 (2014) 236-242.

Ramsay, S., 2011. *Reading machines*, Champaign, IL: University of Illinois Press.

User Required? On the Value of User Research in the Digital Humanities

Max Kemman
University of Luxembourg
max.kemman@uni.lu

Martijn Kleppe
Erasmus University Rotterdam
kleppe@eshcc.eur.nl

Although computational tools play an increasingly important role in the humanities, adoption of tools by scholars does not always reach its potential. (Warwick, Terras, Huntington, & Pappa, 2007). In projects where the research data is published within a tool, this can result in neither the tool nor the research data being used by other scholars. One partial solution to this problem is to publish research data separately from the tool, as advocated by Borgman (2012), and Kansa et al. (2010). In order to create tools that will be adopted by scholars, one approach is user-centred design, which starts with user research to uncover the needs and wishes of the user group, commonly referred to as user requirements. However, it is debatable whether such user requirements can be generalized to a wider group of humanities scholars (Blanke & Hedges, 2013; Unsworth, 2000; van Zundert, 2012), and whether users are able to explicate their requirements for methodological innovation (Nielsen, 2001; Norman, 2010). We ask what the role of user research is within the Digital Humanities. Our research question is: what is the added value of user research for developing tools aimed at digital research methods?

To address this question, we will discuss results from our own user research for gathering user requirements for two Digital Humanities projects; PoliMedia (`http://www.polimedia.nl`) and Oral History Today (`http://zoeken.verteldverleden.org`). In these projects, we held semi-structured interviews with respectively five and fifteen scholars to inform development. We will show how many user requirements were common to multiple participants, and our categorization of the requirements as within- or out-of- scope of the projects' goals.

Our results show scholars have a clear idea how they perform their research, and how tools could simplify steps in the process of discovering and analysing sources. Participants did not limit their needs and wishes to the scope set by the project. First, a large portion of the user requirements was out-of-scope, related to e.g. unavailable metadata or computational processing of sources. Second, only few user requirements were related to the specific technological goals of the projects. Moreover, due to the many unique and out-of-scope user requirements, we note that there is a tension between the specificity of scholarly research methods, and generalizability for a broader applicable tool.

Nevertheless, our findings suggest that user research has a clear benefit for DH projects: first, the user requirements that were within-scope led to usable features that were sufficiently generic for the tool to be adopted for purposes for which it was not specifically created. Second, the out-of-scope user requirements give insight into the tool's compatibility with the individual's wider research workflow. However, this also shows that this wider research workflow cannot be generalized into a single tool.

Therefore, in addition to performing user research for the tool under development, we note that the data should be published independently from the tool. This allows researchers to not only use the data with the tool provided, but also to use required or preferred tools at other steps in their research workflows. By thus combining user research with open data, we expect the tool and the data will be able to reach their full potential.

References

Blanke, T., & Hedges, M. (2013). Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems*, 29(2), 654–661 (http://doi.org/10.1016/j.future.2011.06.006).

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* (http://doi.org/10.1002/asi.22634.

Kansa, E. C., Kansa, S. W., Burton, M. M., & Stankowski, C. (2010). Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies*, 6, 301–326. http://doi.org/10.1007/s11759-010-9146-4

Nielsen, J. (2001). First Rule of Usability? Don't Listen to Users. Retrieved July 1, 2014, from http://www.nngroup.com/articles/first-rule-of-usability-dont-listen-to-users/.

Norman, D. A. (2010). The way I see it: Technology first, needs last. *Interactions*, 17(2), 38 (http://doi.org/10.1145/1699775.1699784).

Unsworth, J. (2000). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? In SymposiumonHumanitiesComputing:FormalMethods,ExperimentalPractice. King's College, London.

Van Zundert, J. (2012). If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities. *Historical Social Research / Historische Sozialforschung*, 37(3), 165–186.

Warwick, C., Terras, M., Huntington, P., & Pappa, N. (2007). If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. *Literary and Linguistic Computing*, 23(1), 85–102 (http://doi.org/10.1093/llc/fqm045).

Using Parallel Data to Improve Part-of-speech Tagging of 17th Century Dutch

Dieuwke Hupkes
University of Amsterdam
dieuwkehupkes@gmail.com

Rens Bod
University of Amsterdam
rens.bod@uva.nl

If one wants to extract information from a (historical) text, it is often useful to know the grammatical categories (or part-of-speech tags) of all the words in this text. Tools to automatically assign high accuracy POS-tags are freely available, but require large amounts of annotated training data. Automatically POS-tagging languages for which little annotated data is available has proved to be a challenging task, and when developing a POS-tagger for historical Dutch one is confronted with an additional difficulty: a large variation in spelling. Currently, digital humanities researchers working with historical Dutch texts often resort to taggers trained on contemporary Dutch (e.g., Van den Bosch et al., 2007), but the accuracy of the resulting tags is generally low.

In this paper, we explore methods for generating higher accuracy tags for historical corpora. In particular, we investigate how information extracted from a diachronic parallel corpus consisting of Dutch Bible texts from 1637 and 1977 can be used to improve POS-tagging of 17th century Dutch texts from several domains. We explore the possibility of using this corpus to rewrite/translate words prior to tagging, as well as the opon of creating an annotated 17th century corpus by projecting tags from the contemporary version of the corpus to its historical counterpart and using this corpus to train a new tagger.

We show that even without applying methods to account for context dependencies, both methods result in great improvements over a baseline of tagging the texts with a tagger trained on contemporary Dutch. Furthermore, they significantly outperform using simple rewrite rules to normalise/modernise spelling before tagging. The improvement subsists across domains, but the within domain results are significantly beer than the results for other domains, suggesting that incorporating knowledge about the domain of the text can lead to further improvement.

The results of this study can be of direct use to digital humanities researchers working with historical Dutch texts of the last 3 centuries, as the tags assigned by our retrained tagger are of much higher quality than the current standard. Furthermore, it shows that using parallel data to exploit the similarities between contemporary and historical texts is a very promising path to developing diachronic taggers. Similar techniques could also be applied to other languages that have diachronic parallel corpora available, as well as to improve results on lemmatisation of historical Dutch texts. In future work, we will focus on domain adaptation techniques for a better consistency across domains.

References

Antal Van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Computational linguistics in the Netherlands: Selected papers from the Seventeenth CLIN Meeting*, pages 99–114, 2007.

Varieties in contemporary Dutch: Combining the research possibilities offered by MIMORE and GrETEL

Liesbeth Augustinus
KULeuven
liesbeth@ccl.kuleuven.be

Ineke Schuurman
KULeuven
ineke@ccl.kuleuven.be

Sjef Barbiers
Meertens Institute KNAW
sjef.barbiers@meertens.knaw.nl

MIMORE and GrETEL are two linguistic search engines. MIMORE enables linguists to query three related databases on morphosyntactic variation in Flemish and Dutch dialects: SAND, DiDDD, and GTRP. The data are based on interviews and elicitation. GrETEL enables linguists to consult syntactically annotated corpora (or treebanks) in a user-friendly way. For Dutch the CGN treebank (spoken Dutch, 1M words),1 Lassy Small (written Dutch, 1M words) [1] and SoNaR (500M words) are available. These data concern actual utterances in the 'standard language'. But is this standard language the same in the entire Dutch-speaking region?

Some questions presenting themselves: Is there any evidence of a correlation with the dialect used in a specific area? Is there a stronger correlation with spoken language than with written language? Thus: can MIMORE be used to interpret findings in GrETEL? Are data found by means of elicitation (MIMORE) reflected in actual standard language use (GrETEL)?

In a combined educative CLARIN-NL use case on verb clusters we investigated those questions. In this case study, we investigated verb clusters of (1) a finite modal verb (e.g. moeten, 'must'), (2) a temporal auxiliary (e.g. hebben, 'have') and (3) a main verb functioning as past participle (e.g. gemaakt, 'made'), as in:

De juf zegt dat hij zijn huiswerk morgen moet hebben gemaakt.
The teacher says that he his homework tomorrow must have made
'The teacher says that he has to make his homework by tomorrow.'

Depending on the dialect the order of the verbs in the sentence may differ, without changing the meaning of the sentence. The first figure below shows the word order variation per region, based on data obtained from MIMORE, while the table below presents the figures of the variation found in GrETEL. In our presentation we will show how we obtained those figures and we will present a series of related findings.

Notes

[1] CGN covers both Dutch as spoken in Flanders (CGN-VL) and Dutch as spoken in the Netherlands (CGN-NL)

References

MIMORE website: http://www.meertens.knaw.nl/mimore/.
GrETEL website: http://gretel.ccl.kuleuven.be.

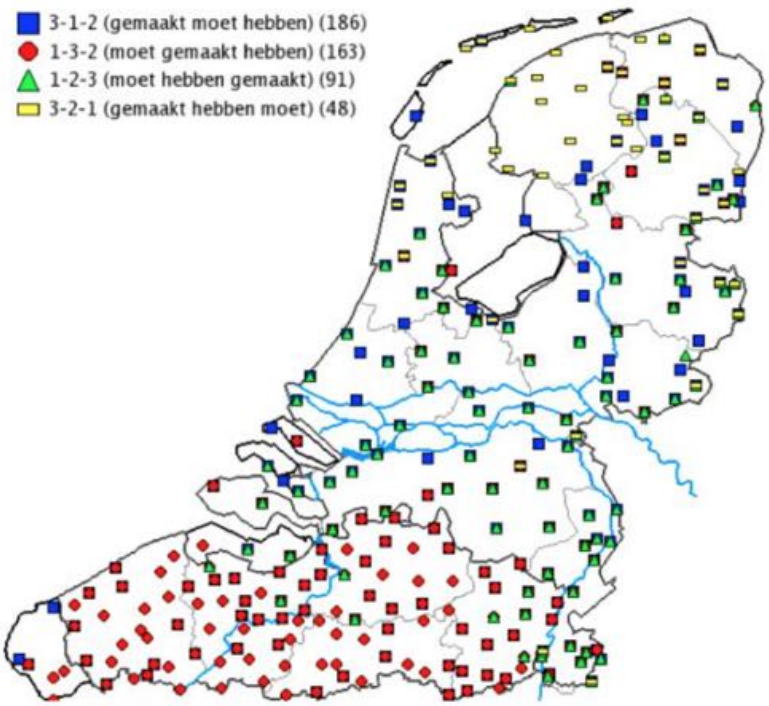


Figure 22: Frequency per region (MIMORE).

.

	CGN (NL + VL)	LASSY	SUM
1-2-3	19 + 7	66	92
1-3-2	0+3	2	5
3-1-2	14+3	33	50
3-2-1	0+0	0	0
2-1-3	0+0	0	0
2-3-1	0+0	0	0
SUM	**33+13**	**101**	**147**

Figure 23: Frequency in the standard language (GrETEL).

Visual Hermeneutics

Peter Verhaar
Leiden University
p.a.f.verhaar@hum.leidenuniv.nl

Conventional literary research is dominated profoundly by the written word, as text frequently forms both the object and the outcome of academic research. This centrality of text is increasingly attenuated, however, as a consequence of a growing use of computational methods. The essential plasticity and computability of digital text has inspired innovative forms of analysis, and the growing interest in quantification and in statistics has urged literary scholars to explore whether or not the salient features of texts can be clarified and communicated effectively though images (Jessop 2008). As studies in the field of literary criticism generally aim to elucidate the meaning of texts, Sinclair et al. argue that the value of tools for the visualisation of texts can be gauged "by determining how well it supports this interpretative activity" (Sinclair et al. 2013). The field of literary studies does not have a historical tradition of using visual displays such as graphs and charts for the organisation and the dissemination of knowledge, however, and many of the visualisation techniques which are in use at present have been adopted from the social sciences and the natural sciences. Johanna Drucker has emphasised that the standardised visualisation models which originated in statistics and in the empirical sciences unfittingly follow the epistemological assumptions of these disciplines, and she urges scholars to develop graphical displays which represent the "observer co-dependent" nature of data and which expose the "interpretative complexity" of humanistic data (Drucker 2011).

In this paper, the results are presented of a study which aimed to explore the capacity of data visualisations to support hermeneutic processes. The study focused, more specifically, on the interpretation of poetry. During an initial phase of the study, software was developed for the recognition of various literary devices, such as rhyme, alliteration, enjambment, onomatopoeia, refrains and imagery. Additionally, a large number of techniques have been developed for the visual representation of these literary phenomena. Graphical displays can be classified in a variety of ways. Visualisations may differ with respect to the type of data they are based on, the type of processing these data have undergone, and the geometric objects that have been used to assemble the graph. Many of the existing visualisation tools represent data about word frequencies, but, as will be shown, the possibilities for interpretation can generally be extended when data about the vocabulary is combined with data about occurrences of specific literary figures. Visualisations may also display data on different levels of analysis. They can be used to expose patterns within corpora in their entirety, but they may also represent data about individual poems. In the latter case, diagrams and charts can support forms of close reading. Minute examinations of the way in which literary devices have been combined at the level of individual stanzas and individual lines can help scholars to analyse the concrete ways in which formal textual phenomena contribute to the meaning of a poem (Chaturvedi et al. 2012).

An important objective of visualisations is to establish a condensation, and to provide a succinct expression of the data that are available. While linear texts are invariably highly complex and multifaceted phenomena, visualisations typically privilege a limited set of dimensions at the expense of certain other dimensions, allowing for a more concentrated investigation of the aspects that remain. Data visualisations are typically created to marshal surprises. They may expose specific arrangements of sounds or of words which are difficult to see during a close reading of the words of the text. The patterns that emerge often stimulate scholars to examine whether or not the groupings that are generated on statistical grounds also coincide with other divisions, such as those based on theme, genre or date of creation. Such attempts to explain unexpected correlations or conspicuous disassociations can in turn galvanise a hermeneutic engagement with the texts that are rendered graphically.

References

Chaturvedi, M. et al. (2012). 'Myopia: A Visualization Tool in Support of Close Reading'. *Digital Humanities Conference 2012*.

Drucker, J. (2011). 'Humanities Approaches to Graphical Display'. *Digital Humanities Quarterly*, 005(1). Available at: http://digitalhumanities.org:8080/dhq/vol/5/1/000091/000091.html.

Jessop, M. (2008). 'Digital visualization as a scholarly activity'. *Literary and Linguistic Computing*, 23(3), pp. 281–293.

Sinclair, S., Ruecker, S. & Milena Radzikowska (2013). 'Information Visualization for Humanities Scholars'. In *Literary Studies in the Digital Age. Modern Language Association of America*. Available at: http://dlsanthology.commons.mla.org/information-visualization-for-humanities-scholars/.

Visualizing medieval book production: Data visualization in medieval manuscript studies

Giulio Menna
Sexy Codicology
giulio.menna@gmail.com

Marjolein de Vos
mjt.de.vos@gmail.com
giulio.menna@gmail.com

Data concerning medieval manuscripts (i.e. place of origin, date of creation, etc.) has been generated thanks to the efforts of many researchers, and it is available, mostly, in printed form. Many of these books have been digitized, but the data, the numbers, contained in them are still difficult to retrieve. The objective of our project is to show and discuss how research results in manuscript studies can be visualized, lead to new and interesting insights, and render data more accessible. We want to argue that approaching this field from the side of digital humanities can be enlightening.

Some of this data is already available from important institutions (Europeana, CERL, Archive.org), but it is often difficult to retrieve for non-computer experts. Furthermore it often overlooks important information from secondary sources (i.e. who dated a manuscript? Where can we find reference material? Who attributed this manuscript to the scribe?). This information is available in printed form, and thanks to OCR technology, it is relatively fast to digitize this data and create datasets. This process transforms "solid data" (data present only in print) into "liquid data" (data available digitally), which is easier to retrieve and mix, while being properly referenced and linked to the original source. Ideally, a visualization would include all primary and secondary information that is available, so that with one click a researcher may have a complete overview of sources. Additionally, incorporating information from other fields of studies such as social studies and history can help contextualizing the information and lead to new insights.

Our example will focus on data concerning a manuscript collection at Leiden University Library. We will create quickly consultable and linked datasets, generate interactive digital maps and infographics, and create a publicly available template to develop comparable end-products with one's own data. The example that will be presented will help showcase the possibilities that are available by data visualization in manuscript studies and related fields. The possibilities in this area have yet to be more fully explored, but they are quite promising.

Visualizing the Dutch Folktale Database

Iwe Everhardus Christiaan Muiser
Twente University
e.c.muiser@utwente.nl

Mariet Theune
Twente University
m.theune@utwente.nl

Humanities researchers are often not well aware of the technical advancements and possibilities offered by computer science. Demand for technical aid is therefore often based on traditional research approaches. Our goal is to bridge the gap between the humanities and computer science by developing tools for the exploration of a folktale collection from a technical perspective. Our intermediate results of these experimental approaches will be presented to folktale researchers to inspire alternative ways of investigation.

In this work we present two visualization approaches based on existing metadata and technologies. Two dimensions of folktales, geographic location and date, are obvious choices for visualization due to their relationship with the historical and geographical environment in which the tales came into existence (Abello et al., 2012). These dimensions can be used to provide a bird's eye view of the collection, or a subset of the collection. Our system supports selecting subsets of the data based on search queries and visualizes them on a map in combination with a timeline (first figure below). A second visualization approach that can be added on top of such collections, is to display similarity between documents. Similarity between folktales is of great interest to researchers who are searching for the origins of tales (Tehrani et al., 2013) however, most investigations of narrative similarity have only been carried out on a small scale so far (Nguyen et al., 2014). In our visualization tool, the similarity of folktales can be determined based on (different configurations of) metadata such as folktale type or keywords, or the text of a document. Our system visualizes the similarity between tales as a dynamic network graph (first figure below).

The folktale collection we use is the Dutch Folktale Database, which was established in 1994 at the Meertens Institute in Amsterdam. Since then it went through several phases of development and now contains nearly 44.000 folktales with standardized and spring cleaned metadata (Muiser et al. 2012). Dynamic visualizations like plotting search results on a map, timeline or network graph could lead to new insights into existing data. Visualizations can also be beneficial to detect mistakes in metadata that were made during annotation, like typing errors, divergent formats, or capitalization/ punctuation errors.

The tools are still in development and will soon be tested to see if previous assumptions based on qualitative research can be verified by modern means of search and visualization. Also, it will be interesting to observe whether integration of the tools in the Dutch Folktale Database website (`www.verhalenbank.nl`) will make it easier for users to browse through the collection. With our work, we hope to open up new avenues for folktale research that go beyond what was previously possible. The experimental visualization interfaces are available online:

1. The Dutch Folktale Map Tool: `http://www.verhalenbank.nl/verhalenkaart/`

2. The Dutch Folktale Network Tool: `http://www.verhalenbank.nl/verhalennetwerk/`

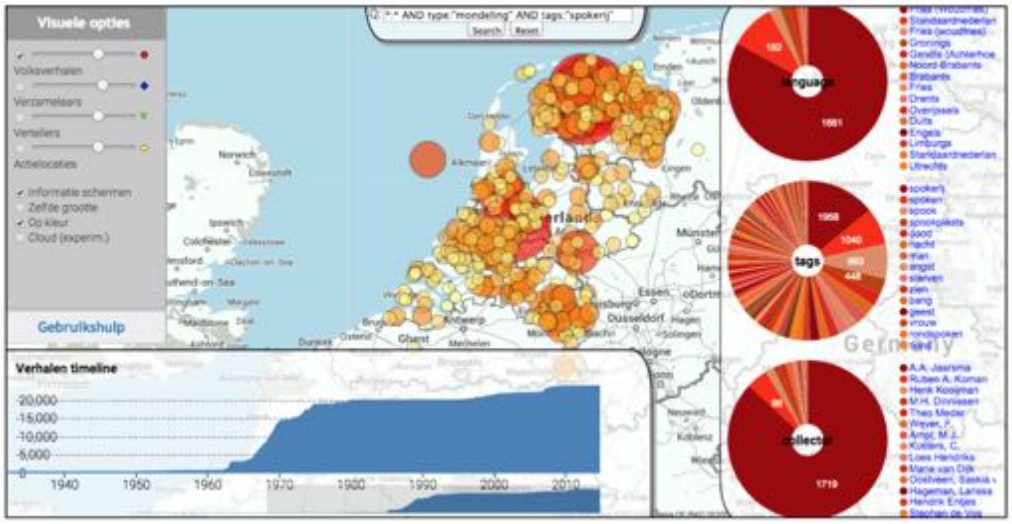


Figure 24: Folktale map browser tool.

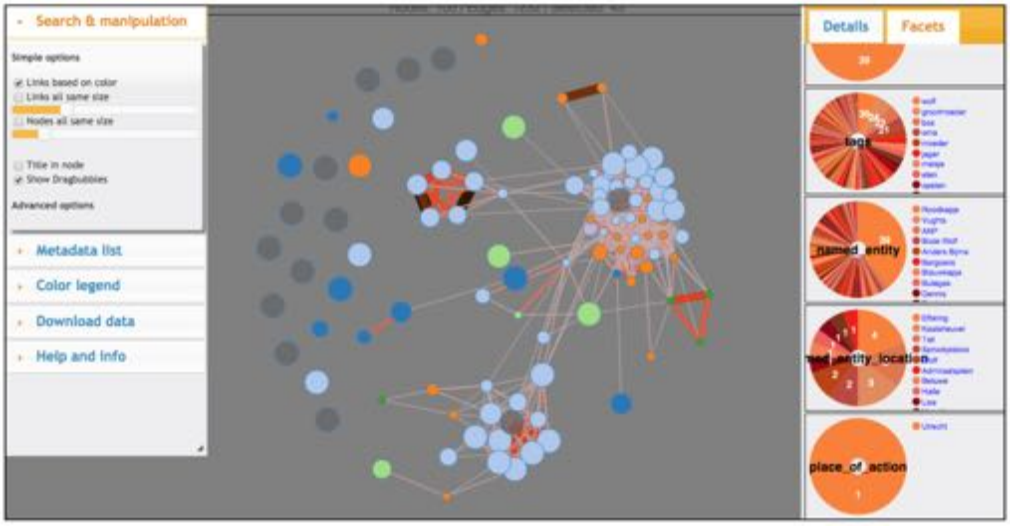


Figure 25: Folktale network browser tool.

References

J. Abello, P. Broadwell, and T. R. Tangherlini (2012) Computational folkloristics. *Communications of the ACM*, vol. 55(7), pp. 60–70.

I. Muiser, M. Theune and T. Meder (2012) Cleaning up and standardizing a folktale corpus for humanities research. *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH2)*, pp. 63–74.

D. Nguyen, D. Trieschnigg and M. Theune (2014) Using Crowdsourcing to Investigate Perception of Narrative Similarity. *Proceedings of CIKM'2014*, pp. 321–330.
J. J. Tehrani (2013) The phylogeny of Little Red Riding Hood. *PloS one*, vol. 8(11):e78871.

Visualizing the Narratives of European Integration

Laurie de Zwart
Radboud University Nijmegen
l.dezwart@student.ru.nl

Sarah Döking
Radboud University Nijmegen
sarah.doking@gmail.com

Nicky van Rijsbergen
Radboud University Nijmegen
n.vanrijsbergen@student.ru.nl

Tijmen Weber
Radboud University Nijmegen
Tijmenweber1@gmail.com

Anna Verhoek
Radboud University Nijmegen
annaverhoek@gmail.com

Zsófi Bognár
Pázmány Péter Catholic University
bogszo88@gmail.com

Adrienn Adolf
Pázmány Péter Catholic University
adolf.adrienn@gmail.com

Réka Köcsky
Pázmány Péter Catholic University
r.kocksky@gmail.com

Zíta Huszthy
Pázmány Péter Catholic University
zita.huszthy@gmail.com

Renata Hrecska
Pázmány Péter Catholic University
dr.h.renata@gmail.com

Iris Hendrickx
Radboud University Nijmegen
iris@i-hx.nl

Antal van den Bosch
Radboud University Nijmegen
a.vandenbosch@let.ru.nl

Zsolt Almási
Pázmány Péter Catholic University
almasi.zsolt@btk.ppke.hu

The study here presented originates from an international student research project organized by the Radboud Honours Academy and IRUN excellence program. We aimed to design a concept for a sustainable and user friendly enhanced publication for cultural heritage institutes in general, but in particular for our client CVCE, the Centre Virtuel de la Connaissance sur l'Europe (centre for European Integration). Here we focus on one of the research questions in our project: in what ways does presentation in a digital environment influence the perception of cultural heritage? Within this question, one of the accents is on visualization techniques as an instrument for narrativity on web pages. The way in which a narrative is told, differs greatly from medium to medium and especially within the digital realm there is a new way in which stories present itself. To clarify, to have an overview at once or to refer to related stories or contexts, visualization techniques such as interactive images, videos, interactive maps and storylines could help to underline and open up the story for a larger audience.

The methodology that we use is the evidence based policy, developed by Pawson and Tilley (2004). This model compares several cases in order to distill a line or conclusion. By analyzing, structuring and comparing several CVCE-like websites on the way that they try to tell a story and on their interpretation and execution of visualization, we develop a concept that helps to rethink the way in which enhanced publications can help scholars in cultural heritage research. We will discuss how media influence storytelling and, in addition, how visualization techniques can be helpful in doing so. Furthermore we will show examples of the weaknesses and strengths of different visualization techniques and present the outcomes of our comparisons.

References

Burgess, Helen J. & Hamming, Jeanne (2011) 'New Media in the Academy: Labor and the Production of Knowledge in Scholarly Multimedia', in: *Digital Humanities Quarterly*, Vol. 5, No. 3.

Giaccardi, Elisa (2012) *Heritage and Social Media: Understanding heritage in a participatory culture*. London/New York: Routledge Ltd.

Hayles, N. K. (2012). *How we think: Digital Media and Contemporary Technogenisis*. Chicago/London, Chicago University Press.

Van den Heuvel, Charles, Hoogwerf, Maarten et. al. (2011) *Dynamic Drawings in Enhanced Publications*. KNAW.

Pawson, Ray & Tilley, Nick (2004) *Realist Evaluation* (`http://www.communitymatters.com.au/RE_chapter.pdf`). (25-02-2015).

What makes dream text dreamy?

Antal van den Bosch
Radboud University
a.vandenbosch@let.ru.nl

Iris Hendrickx
Radboud University
email@authoreo.ne

Maarten van Gompel
Radboud University
proycon@anaproy.c

Ali Hürriyetoğlu
Radboud University
ali.hurriyetoglu@gmail.com

Folgert Karsdorp
Meertens Institute KNAW
Folgert.Karsdorp@meertens.knaw.nl

Florian Kunneman
Radboud University
f.kunneman@let.ru.nl

Louis Onrust
Radboud University
l.onrust@let.ru.nl

Martin Reynaert
Tilburg/Radboud University
reynaert@uvt.nl

Wessel Stoop
Radboud University
a.vandenbosch@let.ru.nl

The analysis of dreams has a long history. One of the earliest recorded dream analyses was written on clay tablets in Mesopotamia, 5000 years ago (Black & Green, 1992). In ancient Greek and Egyptian times, dreams were seen as messages from the gods. Nowadays, many different fields study the meaning and purpose of dreams such as psychiatry, psychology, neuroscience and religious studies, but a definite explanation of the purpose of dreams is still far from being found. Previous studies on content analysis of dreams have shown that the content of dreams reflects a person's daily life and personal concerns. Around 75-80% of dream content relate to everyday settings, characters, and activities. A much smaller part of dreams descriptions related to bizarre topics shared by numerous people like dreaming about flying, teeth falling out or being naked in public (Domhoff & Schneider, 2008).

We aim to detect what it is that makes a dream text different from other texts. What are those features that distinguish a dream description from personal true stories such as personal stories, diary entries or confessions? In our study we investigate several strategies such as n-gram analysis, topic modeling, text classification and discourse coherence measures. We apply these supervised and unsupervised methods to a collection of about 20K dream reports from the benchmark data set Dreambank (www.dreambank.net) and a collected sample of true personal stories. We present the outcomes of our experiments and show to what extent we uncover what makes a dream text dreamy.

References

Black, J., & A. Green (1992). *Gods, Demons and Symbols.* University of Texas Press

Domhoff, G. W., & Schneider, A. (2008). Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition*, 17, 1238-1247

Writing Songs into Literary History with Digital Text Mining

Lisanne Vroomen
Ruusbroec Institute
University of Antwerp
Lisanne.Vroomen@uantwerpen.be

The manuscript Berlin SBB-PK mgo 185 has been written around 1500 and contains 91 devout vernacular songs, which were used by the Sisters of the Common Life in the Dutch city Zwolle. These sisters were part of a larger religious movement, known as the Modern Devotion. In order to position the songs in the textual culture of the Devout, I will compare the content of Berlin 185 with the content of devout prose originating from the same female religious environment. The devout prose that I will use as a reference corpus consists of biographies of the sisters and informal sermons.

Previous research into the textual culture of the Modern Devout has not paid much attention to spiritual songs, a genre that has flourished mainly in female circles (Joldersma 2001, Joldersma 2008 and Van der Poel 2011). Hence, study of song manuscripts can yield new insights into vernacular spiritual literature and the textual culture of Modern Devout. With my research I want to show the unique position of songs within this textual culture.

For the comparison between songs and prose, I use both close-reading and digital methods of texts mining, such as frequency lists, keyword lists and skip grams (WordSmith, Scott 2012; Stylo, Eder e.a. 2014). Several other researchers have shown how these digital methods can shed more light on a specific text, or a specific texts (Arche 2009, Baker 2004, Jockers 2013, Mahlberg 2007 and Stubbs 2005). By combining these methods with close-reading I will be able to get more insight into the texts at a macro level, without losing focus of the micro level.

My hypothesis is that the songs, although they are used in the same environment as the prose, focus on different subjects and have a different function and angle. For example, the songs use the words Mary, Jesus and heaven much more often than the prose, while they give less attention to the concepts of virtues and sins, which are very important in the prose texts. Furthermore, the frequency of certain syntactical categories, such as modal verbs or pronouns, can also inform us about the texts and the way the focus of the genres differences. Want is the most frequent word in the songs, and is also a keyword compared to both the biographies and the sermons. For the sermons, however, shall is the most important modal verb. This is an indication that the songs are more expressive ('I want') than the sermons, which are didactic in nature ('you shall'). The high frequency of the pronoun I in the songs is an indication of the personal character of the songs, and can be contrasted with the use of we in the sermons, which indicates a focus on the community as a whole. This is shown in the figure below. In my presentation I will discuss these examples further and combine them with a close-reading.

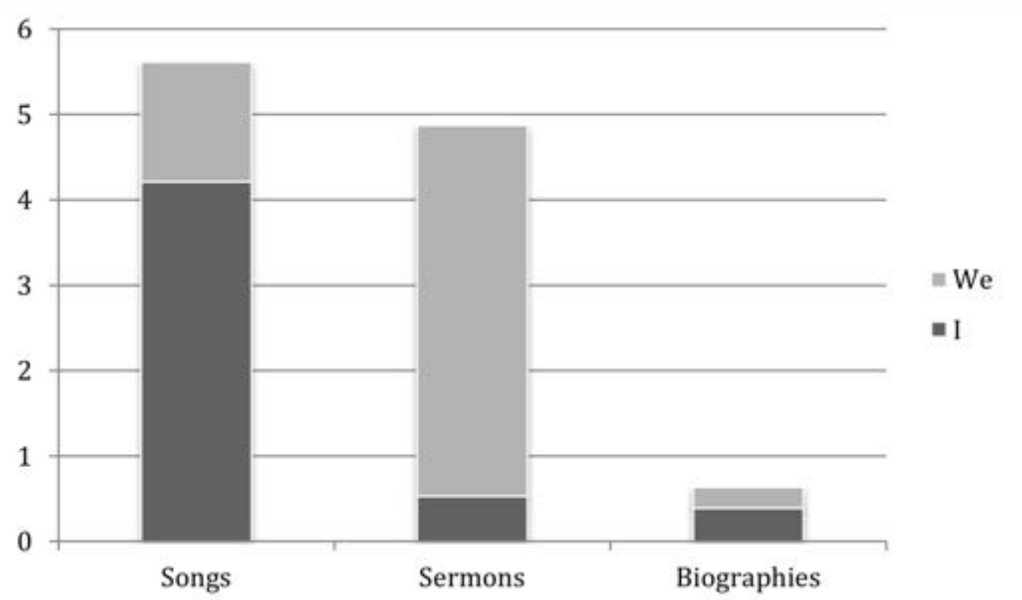


Figure 26: Relative frequency of first person personal pronouns.

References

Arche, D. (2009) *What's in a Word-list? Investigating word frequency and keyword extraction*, Burlington: Ashgate.
Baker, P (2004) 'Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis', *Journal of English Linguistics*, vol. 32, pp. 346–359.
Jockers, M. L. (2013) 'Theme', *Macroanalysis: Digital Methods and Literary History*, Urbana/Chicago etc: University of Illinois Press.
Joldersma, H. (2001) 'Writing Late-Medieval Women and Song into Literary History', *Tijdschrift voor Nederlandse Taal- en Letterkunde*, vol. 17, pp. 5–26.
Joldersma, H. (2008) '"Alternative Spiritual Exercises for Weaker Minds?" Vernacular Religious Song in the Lives of Women of the Devotio Moderna', *Church History and Religious Culture*, vol. 88, pp. 371–393.
Mahlberg, M. (2007) 'Clusters, Key Clusters and Local Textual Functions in Dickens' , *Corpora*, vol. 2, pp. 1-31.
Poel, D. van der. (2011) 'Late-Medieval Devout Song: Repertoire, Manuscripts, Function', *Zeitschrift für deutsche Philologie, Dialog mit den Nachbarn. Mittelniederländische Literatur zwischen dem 12. und 15. Jahrhundert*, pp. 67–79.
Stubbs, M. (2005) 'Conrad in the Computer: Examples of Quantitative Stylistic Methods', *Language and Literature*, vol. 14, pp. 5-24.
Scott, M., (2012) *WordSmith Tools version 6*, Liverpool: Lexical Analysis Software.
Eder, M., Rybicki, J. and Kestemont, M. (2014) 'Stylo: a Package for Stylometric Analyses'.

'Catalogue these books': Digital Editions and the Digital Library

Ronan Crowley
Universität Passau
crowle01@gw.uni-passau.de

With over twenty million volumes contained in Google Books and over ten million in HathiTrust – to take but two examples – LSDI scanning represents the dominant economy of book digitisation. This poster presentation will address the intersections of digital archives with digitised material produced by other means: the digital editions of authors' manuscripts encoded in markup languages (for example, in the TEI XML format). What are the points of contact and friction that obtain between Big Text Data and heavily curated, deeply encoded editions of archival materials? What are the challenges and possibilities for interpretation afforded by competing or complementary scales of digital materiality?

The case study animating the project centres on an historical practice of reading recoverable from the holograph notebooks that James Joyce compiled while writing *Ulysses* (1922), and which are now housed at the National Library of Ireland. Essentially commonplace books, these documents list hundreds of lone words and abbreviated phrases, typically copied verbatim during the course of Joyce's reading. This storehouse provided the raw material for his novel in progress. Assiduous LSDI browsing and coordinated search is recovering the sources so integrated into Ulysses, but to what degree can such material be included in digital editions that prioritise text encoding? What is the compromise, then, for a digital materiality reckoned at scale between, on the one hand, the impracticality of marking-up large-scale data collections and, on the other, the desired granularity of the digital edition?

PRESENTED AS DEMOS ONLY

A Lexicon of Scholarly Editing

Wout Dillen
University of Antwerp
wout.dillen@uantwerpen.be

This demo presentation will offer an interactive demonstration of the website 'Lexicon of Scholarly Editing' (www.uantwerpen.be/lexicon-scholarly-editing), a digital resource that collects definitions of important concepts in the field of Textual Criticism. Based on the Wordpress infrastructure, the author built the Lexicon as part of his work on the 'Creative Undoing and Textual Scholarship (CUTS)' project, funded by the ERC and supervised by Dirk Van Hulle. The project is affiliated to the Centre for Manuscript Genetics (CMG) at the University of Antwerp, and the European Society for Textual Scholarship (ESTS).

Traditionally divided into three main 'schools' (Anglo-American Scholarly Editing, German *Editionswissenschaft*, and French *critique génétique*), the field of Textual Criticism has witnessed a rapprochement of different editorial theories and practices over the last decades, transforming three distinct monolingual discussions into a larger multilingual one. Rather than focusing on the differences between their respective traditions, it is now generally accepted that they each provide a different but equally valuable perspective on how the texts of literary works could be edited, based on the historical evidence of textual transmission. However, even though more and more theoretical works have been translated outside of their original language, and the communication and mutual understanding between textual critics around the world has never been better, the Lexicon may demonstrate that some difficulties with the more nuanced meanings of theoretical concepts still persist.

The intent of the Lexicon is not to write new definitions for theoretical concepts, but rather to gather existing definitions, and to let those definitions speak for themselves. To achieve this goal, quoted definitions of each concept are ordered chronologically, and rendered in their original language. As such, the Lexicon aims to display these definitions as if in a multilingual discussion with one another – which, in some cases, they quite literally are. By confronting the user with a wide array of different views on theoretical concepts, the Lexicon will allow its users to place them in their proper context before using them in their own work.

The Lexicon was published online on 19 November 2013, and is freely available for anyone who is interested in Textual Criticism. As a work in progress, everyone is invited to collaborate on the Lexicon, either by sending in new definitions (as a registered user, or via a contact form), or by suggesting articles to be added to the Lexicon's To-Do list on the project's public Zotero group (www.zotero.org/groups/lexicon_of_scholarly_editing). The Lexicon currently has three official contributors: Frederike Neuber (a DiXiT fellow based at the Centre for Information Modelling – Austrian Centre for Digital Humanities at the University of Graz, Austria); Elisa Nury, (a PhD student in Digital Humanities at King's College London, UK); and Elli Bleeker, (a DiXiT fellow based at CMG at the University of Antwerp).

A Preview of CENDARI. A Digital Research Infrastructure

Stijn van Rossem
Cendari Project Officer
CERL
stijn.vanrossem@cerl.org

The Cendari project stands for Collaborative European Digital Archive Infrastructure (CENDARI). It is developed by a consortium of 14 partners from 7 countries and is funded by the European Commission as part of the 7th Framework Programme. Cendari is a research infrastructure project aimed at integrating digital resources from the medieval and World War One eras, with an emphasis on those that might be less well known.

Cendari incorporates archival data and creates a research space where users can see projects through from finding and organising sources, to analysing and sharing data with sophisticated tools. The project overcomes the national 'silos' of digitisation efforts and historical inquiry. Perhaps above all it may help open digital history to the majority of professional historians, representing a major breakthrough in digital cultural empowerment.

The aim of CENDARI is to produce a research space (the VRE, or Virtual Research Environment) which will allow scholars to access historical resources across institutional and national boundaries. The core of the VRE is the Note Take Environment (NTE), along with the Archival Directory, and a search functionality (standard, faceted and semantic search). The NTE will allow historians, archivists and librarians to create and upload documents and take notes. In the NTE, the user can compile and comment on sources from the Archival Directory. Historians will also be able to analyze data with the help of sophisticated data mining and visualisation tools. They can submit any text, files or images to Named Entity Recognition, so that the documents can be annotated semantically and connected with controlled vocabularies and ontologies. The user will be able to put their compilation in order, enhance their search for archival material and share and export the project.

In the NTE, the user will be able to access the Archival Research Guides (ARGs). These guides are aimed at historians who are beginning to research a certain topic and assist them in accessing suitable documentation. The Archival Research Guides are based on transnational themes, using archival collections from different countries. CENDARI will create twenty-five freely accessible guides. In addition, users will be able to annotate the existing Archival Research Guides, along with creating their own guides with the Note Taking Environment.

Finally the Archival Directory is a large database of archival collections that will be integrated onto the VRE. It is based on collections relating to CENDARI's two case studies: the Middle Ages and World War One. The user can search and browse through the Archive Directory, which includes both well-known collections and smaller local archives that have been underused by researchers. In addition, scholars can edit and add archival descriptions to the Archival Directory.

References

www.cendari.eu

Bridging Knowledge Collections: Integrating the museum and library systems at the Royal Museums of Art and History (RMAH), Brussels

Ellen van Keer
RMAH – KULeuven
evankeer@vub.ac.be

The Royal Museums of Art and History in Brussels (*www.kmkg-mrah.be*) are one of the 10 Federal Scientific Institutions (FSI) and among the largest museums of art and archaeology in Belgium. Thousands of art treasures and historical objects from around the globe and dating to all periods of human history are on display or kept in the storerooms. Besides its public and educational involvement, the museum engages in scientific research. A range of museum publications and a specialized research library support these activities.

In the last decade, the Belgian Science policy (*www.belspo.be*) has initiated a major digitization campaign in order to improve the management and the public and scientific exploitation of the Federal collections in a digital way (Mettens 2011). As an important result, the RMAH is using dedicated management software today. The objects in its ownership are catalogued in a collection management system (MuseumPlus by Zetcom), published on the museum's online collections website (`www.carmentis.be`) and harvested through an international domain aggregator for museums (www.europeana.eu). Library materials, on their part, are catalogued in an integrated library system (Alma by Ex-Libris/Libis), made accessible online through the library's OPAC (`www.limo.be`), and exported to a union catalog (`www.unicat.be`). Hence the RMAH is producing qualitative and interoperable museum and library metadata.

Nevertheless, as traditionally, the museum and library systems have been operating completely independently so far. They make use of a separate set of domain-specific metadata standards that create "silos" of information (Ellings & Waibel 2007). However, there is a fundamental semantic overlap between the materials and content contained in them. Moreover, this overlap is not limited to the general thematic level (e.g. books and objects in relation with ancient Egyptian culture) but goes all the way down to the individual item-level (e.g. object X is discussed on page Y in book Z). This is especially so in the case of museum libraries which have the specific task of collecting the (scientific) documentation on the museum objects.

In line with the growing movement towards convergence in the heritage sector (Zorich e.a. 2008), achieved particularly in a digital way (Erway & Prescott 2010), the project "Bridging Knowledge Collections" was designed to build comprehensive integrations between the existing management systems with related content at the RMAH. More specifically, the project has investigated and implemented a two-way scenario linking objects in the museum system and documents in the libary system (see figures below). Two new tools were introduced, one for the input of scholarly object bibliography in the museum system that links to corresponding records in the library system, and one for harvesting, indexing and enriching document records with linked object

records in the library system. The pilot focused on the Federal glyptic collections, which allowed further enrichment with newly created materials in other projects, especially full-text objects (Orfeo) and 3D images (Glypcol). Hence, management systems can evolve into research tools.

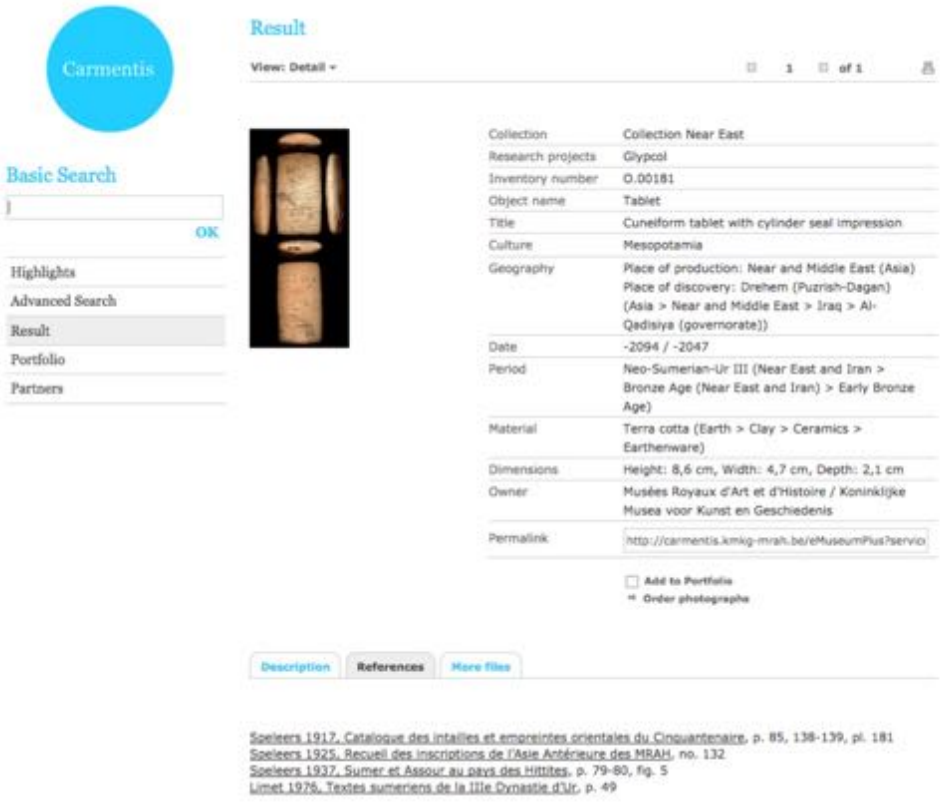


Figure 27: Record of a cuneiform tablet in the online museum catalog Carmentis, enriched with related bibliography.

References

Ellings, M. and Waibel, G. (2007) "Metadata for All: Descriptive Standards and Metadata Sharing across Libraries, Archives and Museums", *First Monday* 12/3 (`http://firstmonday.org/issues/issue12_3/elings/index.html`).

Prescott, L. and Erway, R. (2010) "Single Search: the quest for the holy grail", OCLC Research Report ((http://www.oclc.org/research/publications/library/2011/2011-17.pdf).

Mettens, Ph. (2011) "Het digitaliseringsplan voor de Federale Wetenschappelijke Instellingen en het Koninklijk Belgisch Filmarchief", *ABB* 82, p. 15–30.

Zorich, D., Waibel, G. and Erway, R. (2008) "Beyond the Silos of the LAMs: Collaboration Among Libraries, Archives and Museums", OCLC Research report (`www.oclc.org/research/publications/library/2008/2008-05.pdf`).

Figure 28: Record of a publication in the online library catalog LIMO, enriched with related museum objects.

Ebacs: a Minimalistic Conference Manager

Chris Emmery
University of Antwerp
chris.emmery@uantwerpen.be

Many existing (online) tools for hosting conferences offer a wide variety of functions for large conferences that undoubtedly require these. However, for small conferences that do not need such an advanced system, configuring these systems can be a daunting task. A static conference web page, a small panel of reviewers, and some manual labour to deal with management and finances are, more often than not, sufficient. In this case, a minimalistic service with the exact purpose of making sure the submissions get from the researchers into conference proceedings would keep effort in accordance with the size of the conference. Tying in with this idea, I will present a concept version of an open-source service for conference submission and reviewing. It intends to offer a very minimalistic online experience in setting up conferences, programmes, adding reviewers and handling reviews, and sending e-mail notifications. Additionally, it will allow for directly editing and exporting the conference proceedings to LaTeX, ready for printing. By making it open-source, the hope is that the community might eventually add to the system in line with their own wishes. Potential additions might include utilizing the modular set-up of the system to add custom LaTeX templates, improve preprocessing of submissions, easy integration of more complex submission content, and automatic handling of reference lists. Adding to the effectiveness, and not so much to the complexity of the environment is essential. As such, the time spent on performing and combining most of the small tasks more humble conferences require is kept to a minimum, and most importantly: knowledge of the system's back-end is surely helpful for the community, but certainly not required for your conference.

Many existing (online) tools for hosting conferences offer a wide variety of functions for large conferences that undoubtedly require these. However, for small conferences that do not need such an advanced system, configuring these systems can be a daunting task. A static conference web page, a small panel of reviewers, and some manual labour to deal with management and finances are, more often than not, sufficient. In this case, a minimalistic service with the exact purpose of making sure the submissions get from the researchers into conference proceedings would keep effort in accordance with the size of the conference.

Tying in with this idea, I will present a concept version of an open-source service for conference submission and reviewing. It intends to offer a very minimalistic online experience in setting up conferences, programmes, adding reviewers and handling reviews, and sending e-mail notifications. Additionally, it will allow for directly editing and exporting the conference proceedings to LaTeX, ready for printing.

By making it open-source, the hope is that the community might eventually add to the system in line with their own wishes. Potential additions might include utilizing the modular set-up of the system to add custom LaTeX templates, improve preprocessing of submissions, easy integration of more complex submission content, and automatic handling of reference lists. Adding to the effectiveness, and not so much to

the complexity of the environment is essential. As such, the time spent on performing and combining most of the small tasks more humble conferences require is kept to a minimum, and most importantly: knowledge of the system's back-end is surely helpful for the community, but certainly not required for your conference.

Iter Community: Enabling Social Bibliography and User Project Creation

Shawn DeWolfe
University of Victoria
sdewolfe@uivc.ca

Daniel Sondheim
University of Victoria
sondheim@uvic.ca

Matthew Hiebert
University of Victoria
hiebert8@uvic.ca

William Bowen
University of Toronto Scarborough
william.bowen@utoronto.ca

Ray Siemens
ETCL University of Victoria
siemens@uvic.ca

The Iter community sought to facilitate and support communication, collaboration, and digital project creation for research communities of the Middle Ages and Renaissance [1]. In 2008, the Iter Community was envisioned as a social knowledge creation environment a Community space for collaboration, social networking and the hosting of Iter projects [2]. Now, a team made up of programmers and researchers from the University of Victoria and the University of Toronto, working with an advisory committee, are developing Iter Community in response to community needs [3]. It's a space for social conferences, editions, and bibliographies, as well as for researchers at all career stages to share their scholarly activities with each other.

Furthermore, Iter Community has the following affordances:

1. Tying Iter Community to the core Iter Bibliography 1.3 million record set;
2. Providing metadata for points of interrelation from bibliography to community;
3. Exploring social interactions with record sets (not just reading works cited, but investigating reading practices in the bibliography);
4. Maintaining discussions over user created records that may have been put on hold;
5. Promoting user capability to work with instantly citable records from libraries they curate.
6. Work was undertaken by the team at the University of Victoria's ETCL to create two key elements of the Iter Community project:
   (a) Iter Community Press : a citation plugin that pulls from Zotero records as well as a local database of MARC records to create libraries of bibliographic records that users add to their documents. By combining the data available from Iter Gateway along with the personal citation collections of Iter Community members, users can access a large repository of records to cite in their submissions to the Iter Community network.

(b) The Sandboxer : a WordPress plugin to allow collaborators to deploy an empty project space in Drupal or WordPress that would integrate and feed its project updates back to the Iter Community site.

The Iter Community project builds on WordPress's "Commons In A Box" with custom plugins added to fulfill the goals of the Iter Community project. Our demonstration will showcase how Iter Community lets site users create their own projects and interoperate with the community site, and give project owners the autonomy to turn their sandbox into their own full fledged projects.

Notes

[1] For a history of Iter, see Castell (1997); Bowen (2000); and Bowen (2008).

[2] See the vision document written by Iter Associate Director Ray Siemens (2008).

[3] Needs include "a community's 'scholarly primitives,' . . . from the perspective of professional/user interactions" (Ibid.).

References

Bowen, W.R. (2000, Special Issue). 'Iter. Where does the path lead?', *Early Modern Literary Studies*, 5.3(4), 2126.

Bowen, W.R. (2008). 'Iter: Building an effective knowledge base?', In W.R. Bowen & R.G. Siemens (Eds.), Renaissance studies and new technologies (pp. 101-109).

Castell, T. (1997). 'Maintaining web based bibliographies: A case study of Iter, the bibliography of Renaissance Europe'.

Siemens, Ray. Initial Steps, Following a Larger Vision: A Feature Oriented Pilot Proposal. 2008.

Nederlab, online laboratory for humanities research on Dutch text collections

Hennie Brugman
Meertens Institute KNAW
hennie.brugman@meertens.knaw.nl

March 2015, Nederlab released the first version of its research portal. The Nederlab project (www.nederlab.nl) collects, harmonizes, curates and interlinks all digitized Dutch texts that appeared in print from about 800 until now and makes them available for digital humanities scholars through a virtual research environment. The focus of this VRE is on detection of patterns of change in the Dutch language and culture. Nederlab is built by a team of editors, technicians and (NLP) tools experts at the Meertens Institute, Huygens ING, INL and several Dutch universities and is funded by NWO, KNAW, CLARIAH and CLARIN-NL.

The current Nederlab release is the result of the foundational work done in the first years of the project and is the starting point for gradual extension of the research portal with analytic tools. We carefully designed a metadata schema that is fine-tuned for Nederlab purposes and selected a basic format for Nederlab text content. We implemented a (meta-)data preprocessing pipeline and interactive tools to support our editorial staff. We tested and fine-tuned these by integrating three collections: digitale bibliotheek voor de Nederlandse letteren (DBNL), Early Dutch Books Online (EDBO) and Dutch newspapers until 1900 (National Library of the Netherlands). We designed and implemented indexes and a search web service, to enable fast and large-scale search and refinement on basis of any combination of metadata and text criteria, in any order. Currently, we host 13.5 million titles, ranging from short newspaper articles to long books.

The Nederlab research portal is open to everyone for searching. Moreover, once logged in users enter their personal workspace. This is where they can keep their own research collections and objects, and from where they can start an increasing number of analytic tools. One of the first tools available is a visualisation service that shows distributions of document counts over a number of metadata dimensions, including distributions of document counts over time. The Nederlab project ends at the beginning of 2018. Until then we expect to integrate many new collections at an increasing pace. Nederlab development will be more and more driven by use cases of scientific end users. This will result in well tested analytic tools that have proven to be useful for historians, literary scholars as well as linguists.

The Beckett Digital Manuscript Project and Beckett's Personal Library

Dirk van Hulle
University of Antwerp
dirk.vanhulle@uantwerpen.be

Vincent Neyt
University of Antwerp
vincent.neyt@uantwerpen.be

The Beckett Digital Manuscript Project (www.beckettarchive.org) is a collaboration between the Centre for Manuscript Genetics (University of Antwerp), the Beckett International Foundation (University of Reading) and the Harry Ransom Humanities Research Center (University of Texas at Austin), with the kind permission of the Estate of Samuel Beckett. The Beckett Digital Manuscript Project consists of two parts:

- (a) a digital archive of Samuel Beckett's manuscripts, organized in 26 research modules. Each of these modules comprises digital facsimiles and transcriptions of all the extant manuscripts pertaining to an individual text, or in the case of shorter texts, a group of texts.
- (b) a series of 26 volumes, analyzing the genesis of the texts contained in the corresponding modules.

The Beckett Digital Manuscript Project aims to contribute to the study of Beckett's works in various ways: by enabling readers to discover new documents and see how the dispersed manuscripts of different holding libraries interrelate within the context of a work's genesis in its entirety; by increasing the accessibility of the manuscripts with searchable transcriptions in an updatable digital archive; by highlighting the interpretive relevance of intertextual references that can be found in the manuscripts. The Project may also enhance the preservation of the physical documents as users will be able to work with digital facsimiles.

The purpose of the Beckett Digital Manuscript Project is to reunite the manuscripts of Samuel Beckett's works in a digital way, and to facilitate genetic research: the project brings together digital facsimiles of documents that are now preserved in different holding libraries, and adds transcriptions of Beckett's manuscripts, tools for bilingual and genetic version comparison, a search engine, and an analysis of the textual genesis of his works. The work on this project proceeds in a modular way. Once the electronic genetic edition of a work is completed, the accompanying analysis of the work's genesis is published in print with a selection of facsimile images.

This demo presentation offers an interactive demonstration of the Beckett Digital Manuscript Project, focusing on its latest module: Samuel Beckett's Personal Library, which traces the writing process of Beckett's works back to the source texts in his (extant and personal) library.

References

Beckett, Samuel. (2011) *Stirrings Still / Soubresauts and Comment Dire / what is the word: An Electronic Genetic Edition* (Series 'The Beckett Digital Manuscript Project' module 1), edited by Dirk Van Hulle and Vincent Neyt. Brussels, University Press Antwerp (ASP/UPA). http://www.beckettarchive.org (accessed on 3 April 2015).

Beckett, Samuel. (2013) *L'Innommable / The Unnamable: An Electronic Genetic Edition* (Series 'The Beckett Digital Manuscript Project', module 2), edited by Dirk Van Hulle, Shane Weller and Vincent Neyt. Brussels, University Press Antwerp (ASP/UPA). http://www.beckettarchive.org (accessed on 3 April 2015).

PRESENTED AS POSTERS ONLY

DARIAH and the Benelux

Sally Chambers
Ghent University, DARIAH-BE
sally.chambers@ugent.be

Maarten Hoogerwerf
DANS, DARIAH-NL
maarten.hoogerwerf@dans.knaw.nl

Jan van der West
DANS, DARIAH-NL
jan.van.der.west@dans.knaw.nl

Marianne Backes
CVCE, DARIAH-LU
marianne.backes@cvce.eu

DARIAH, the Digital Research Infrastructure for the Arts and Humanities, aims to enhance and support digitally-enabled research and teaching across the humanities and arts. By bringing together national activities from Member countries, DARIAH is able to offer a portfolio of services and activities centred around research communities. DARIAH was established as a European legal entity in August 2014 with 15 countries – Austria, Belgium, Croatia, Cyprus, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, The Netherlands, Slovenia and Serbia – as Founding Members. This was an important step towards developing a research infrastructure for sharing and sustaining digital arts and humanities knowledge across Europe and beyond.

Strengthening cooperation in the Benelux

Using the DH Benelux conference as a catalyst, DARIAH-BE, DARIAH-LU and DARIAH-NL would like to explore how closer collaboration can strengthen their participation in DARIAH both within their individual countries and together as the Benelux region. Here is an overview of the DARIAH activities in each country:

DARIAH in Belgium

DARIAH-BE is represented by three consortia: the research groups and universities of DARIAH-Flanders (DARIAH-VL) and DARIAH-Federation Wallonia-Brussels (DARIAH-FWB), together with the federal scientific institutions, e.g. Royal Institute for Cultural Heritage and CEGESOMA, the Centre for Historical Research and Documentation on War and Contemporary Society (DARIAH-FED). The key objectives are to identify and join-up existing digital humanities activities across Belgium in order to create a sustainable network of digital humanities researchers that operates within a larger European network. DARIAH-BE will develop and offer a flexible and standards-based service platform consisting of modules and applications for collaborative research, analysis and publishing.

DARIAH in Luxembourg

DARIAH-LU is still in its early stages of development. The current group of partners includes a variety of organisations which reflect the Luxembourg public research landscape and their stakeholders including the Centre Virtuel de la

Connaissance sur l'Europe (CVCE) as the DARIAH National Coordinating Institution for Luxembourg. Luxembourg's participation in DARIAH-EU will build on the strengths of the different partner institutions with a particular focus on the provision of expertise and services/tools for data curation, enrichment, analysis, visualisation and personalisation (including multilingual, multimedia and spatial dimensions).

DARIAH in The Netherlands

In the Netherlands, DARIAH and CLARIN have joined forces in CLARIAH, which brings together the major humanities research institutes in The Netherlands. CLARIAH, which kicked off in March 2015, is a four year roadmap project that aims to provide a common digital infrastructure for the humanities in the Netherlands, focusing on three pillars: media studies, linguistics and socio-economic history. The resulting infrastructure allows ground-breaking research by providing humanities scholars with access to a large collection of digital resources and innovative processing tools.

DARIAH: an opportunity for the Benelux?

Using the opportunity to present a poster at DH Benelux 2015 as a starting point, the authors would like to explore how the three DARIAH countries could collaborate. Initial ideas include:

- increasing collaboration between researchers and infrastructure providers: taking advantage of the geographical proximity and language synergies to participate in shared activities e.g. joint research projects and training events.
- increasing funding opportunities: exploring regional possibilities for funding and establishing partnerships for European funding proposals.
- sharing DARIAH knowledge and experience: increasing understanding and identifying synergies between the DARIAH activities in each country.

Through strengthening the collaboration between DARIAH activities in Belgium, Luxembourg and The Netherlands, we would like to facilitate maximum participation of digital humanities researchers in the Benelux region in DARIAH in order to take full advantage of the benefits of being part of the European DARIAH community.

References

DARIAH-EU: www.dariah.eu.

DARIAH-BE: http://be.dariah.eu/andwww.ghentcdh.ugent.be/content/dariah-be.

DARIAH-LU: http://dariah.lu/andwww.cvce.eu/en/actualites/2014/viewer_news/-/news/55986?refererPlid=10184.

DARIAH-NL: see CLARIAH:www.clariah.nlandwww.clariah.nl/proposal/CLARIAH_2013_Final.pdf.

DiXiT – Digital Scholarly Editions Initial Training Network

Elli Bleeker
University of Antwerp, DiXiT
elli.bleeker@uantwerpen.be

Aodhán Kelly
University of Antwerp, DiXiT
aodhan.kelly@uantwerpen.be

Scholarly editors were quick to realise the value of using digital technology for performing research, disseminating results, facilitating access and bringing research communities together. The early adoption of digital technologies has led to some interesting digital editions, but this initial head start also created some 'handicaps' or challenges for the field. To name but a few: the continuous flux of technological developments requires editors to constantly adapt their tools and methods, and confronts them with serious issues regarding long-term access and preservation. Moreover, the traditional isolated nature of scholarly editing resulted in a range of individual projects reinventing the digital wheel and developing a separate methodology. Another challenge can be found in the designated audience of a digital edition. A digital edition has potential to be of interest to a large audience. However, the public of scholarly editions is traditionally quite small, specific and mainly academic. Finally, while scholarly editing projects are often used as an illustration of the multidisciplinary character of humanities research, their collaborative potential is yet to be fully realised.

This poster presents the DiXiT-project, a European Marie Curie funded research network designed to study the theory, concepts and practices of digital scholarly editing. By exploring the relationship between traditional forms of editing and new computational methods and technologies, DiXiT aims to determine the essential components of scholarly editing in the digital realm. In order to accomplish this, DiXiT distinguished several critical issues, grouped together in specialised work packages. The projects in work package 1 concentrate on concepts, theory, and practice. Work package 2 holds a number of projects that study the technology, standards, and software. Finally, the third work package focuses on the place of scholarly editing within academia, cultural heritage and society as a whole. A number of digital editions will be created within the DiXiT framework, which illustrates how theory and practice are closely intertwined. Consequently, the DiXiT network operates on matters across the full editorial cycle: from digitisation and production through dissemination.

The project runs from 2013 until 2017 and offers a coordinated training and research programme for 12 early stage researchers and 5 experienced researchers in the multi-disciplinary skills, technologies, theories, and methods of digital scholarly editing. Currently 14 of these positions are taken. The 11 early stage researchers are (in alphabetical order) Francisco Javier Álvarez Carbajal, Elli Bleeker, Misha Broughton, Federico Caria, Richard Hadden, Aodhán Kelly, Merisa Martinez, Frederike Neuber, Daniel Powell, Elena Spadini and Tuomo Toljamo. The 3 experienced researchers are (in alphabetical order) Gioele Barabucci, Linda

Spinazze and Magdalena Turska. As a group of DiXiT-fellows we would like to present the current stage of the project and its participants, as well as outline the plans of our specific research projects for the coming years. During the poster presentation, we highlight a number of specific issues from different research projects within each work package. As such, we can provide a more detailed overview of the network's progress as well as a reflection of the current challenges of the field.

Digital Scholarship at the Koninklijke Bibliotheek

Lotte Wilms
Koninklijke Bibliotheek
lotte.wilms@kb.nl

Steven Claeyssens
Koninklijke Bibliotheek
steven.claeyssens@kb.nl

The National Library of the Netherlands (KB) is an active partner in national and international cooperative efforts to develop new knowledge and technology. With this poster, we wish to showcase what the KB can offer researchers in the field of Digital Humanities. We will present our digitised data sets, current research projects and the services of the KB Research Lab that was launched at DHBenelux 2014. The KB has planned to have digitised and OCRed its entire collection of books, periodicals and newspapers from 1470 onward by the year 2030. Already in 2013, 10% of this enormous task was completed, resulting in 73 million digitised pages, either from the KB itself or via public-private partnerships as Google Books and ProQuest. Over 1 million books, newspapers and magazines are currently available via the search portal www.delpher.nl. Next to this, most of these data sets are freely available for research purposes and we welcome and encourage experiments and new applications. The virtual KB Research Lab shows some of such applications and invites researchers to experiment with our data, new technologies and innovative prototypes. The KB also collaborates with researchers in research projects or (junior) fellowships to learn from their research in order to improve the services we provide for Digital Humanists. This poster will present the various data sets that the KB has available for research, the activities we undertake to work together with scholars in research projects and the services that we offer those who wish to work with our material.