Computational Literary Studies Infrastructure (CLS INFRA): Initial Findings and Conclusions for the Field

Birkholz, Julie M.
julie.birkholz@ugent.be
Universiteit Gent

Börner, Ingo
ingoboerner86@gmail.com
Universität Potsdam

Byszuk, Joanna
joanna.byszuk@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Chambers, Sally
sally.chambers@ugent.be
Universiteit Gent

Charvat, Vera Maria
veramaria.charvat@oeaw.ac.at
Austrian Academy of Sciences, Austrian Centre for Digital Humanities and Cultural Heritage

Cinková, Silvie
cinkova@ufal.mff.cuni.cz
Charles University, Prague

Dejaeghere, Tess
tess.dejaeghere@ugent.be
Universiteit Gent

Dudar, Julia
dudar@uni-trier.de
Universität Trier

Ďurčo, Matej
matej.durco@oeaw.ac.at
Austrian Academy of Sciences, Austrian Centre for Digital Humanities and Cultural Heritage

Eder, Maciej
maciejeder@gmail.com
Institute of Polish Language (Polish Academy of Sciences)

Edmond, Jennifer
jennifer.edmond@dariah.eu
Trinity College Dublin

Fileva, Evgeniia
fileva@uni-trier.de
Universität Trier

Fischer, Frank
fr.fischer@fu-berlin.de
Universität Potsdam

Garnett, Vicky
vicky.garnett@dariah.eu
DARIAH-EU

Heiden, Serge
slh@ens-lyon.fr
Ecole Normale Supérieure, Lyon

Křen, Michal
michal.kren@ff.cuni.cz
Charles University, Prague

Kunda, Bartłomiej
bartlomiej.kunda@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Laszakovits, Sabine
sabine.laszakovits@oeaw.ac.at
Austrian Academy of Sciences, Austrian Centre for Digital Humanities and Cultural Heritage

Mrugalski, Michał
michal.mrugalski@hu-berlin.de
Humboldt-Universität zu Berlin

Papaki, Eliza
eliza.papaki@dariah.eu
DARIAH-EU

Raciti, Marco
marco.raciti@dariah.eu
DARIAH-EU

Resch, Stefan
stefan.resch@oeaw.ac.at
Austrian Academy of Sciences, Austrian Centre for Digital Humanities and Cultural Heritage

Ros, Salvador
sros@scc.uned.es
UNED Madrid

Schöch, Christof
schoech@uni-trier.de
Universität Trier

Šeļa, Artjoms
atrjoms.sela@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Tasovac, Toma
ttasovac@humanistika.org
Belgrade Center for Digital Humanities

Tonra, Justin
justin.tonra@nuigalway.ie
University of Galway

Tóth-Czifra, Erzsébet
erzsebet.toth-czifra@dariah.eu
DARIAH-EU

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam

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

van Rossum, Lisanne
lisanne.van.rossum@huygens.knaw.nl
Huygens Institute

Introduction

The Computational Literary Studies Infrastructure (CLS INFRA) project aims to survey and federate resources currently available in digital libraries, archives, repositories, websites or catalogues, with the tools needed to interrogate them, and with a widened base of users, in the spirit of the FAIR and CARE principles (Wilkinson et al. 2016, Carroll 2020), and to address a lack of standardisation that hinders how they are constructed, accessed and the extent to which they are reusable (Ciotti 2014). The aim is not to build entirely new resources, but rather build on recently-compiled high-quality literary corpora, such as DraCor and ELTeC (Fischer et al. 2019, Burnard et al. 2021, Schöch et al. to appear), and to integrate existing tools for text analysis, e.g. TXM, stylo, multilingual NLP pipelines (Heiden 2010, Eder et al. 2016), taking advantage of deep integration with two other infrastructural projects, namely the CLARIN and DARIAH ERICs.
In the proposed poster we want to present achievements of the CLS INFRA projects in its duration so far, focusing specifically on four recent deliverables.

Baseline Methodological User Needs Analysis

In order to document shared research practices and provide an empirical basis for the definition of training needs and infrastructural requirements, we have built a corpus of CLS publications published in 2010-2021. We have then identified the frequency with which specific (a) tools and software, (b) data formats, and (c) methods of analysis are mentioned in this corpus, as well as what are the most common formats, tools and methods. We examined the development of their mentions over time, and attempted to explain the quantitative results.
The key outcomes of the study consist of the following elements: (1) A corpus of research articles, with several metadata tables that includes key information on the collected publications, and containing publications marked as DH more generally, or to CLS more specifically; (2) A dataset and a collection of visualizations that provide information about the frequency and distribution of mentions of tools, formats and methods in the corpus mentioned above (see chapter 3); (3) A report that outlines the composition of the corpus used, explains the methodological steps undertaken, summarizes the key findings based on the data and derives conclusions from these findings.

Skills Gap Analysis

Here we explored current gaps in teaching of research skills for computational literary studies, to adjust own approach to training schools, but also chart the territory for broader communities of practitioners and scholars. We approached the task primarily in a quantitative manner to be able to uniformly bring major tendencies in teaching and skill demand to a common denominator. This required an explicit mapping of (1) existing teaching practices (“supply”) and (2) opinions of the practitioner community (“demand”) to a single grid of skills, where it is possible to compare both parts and identify the gaps directly. Skills in the grid were derived from four broadly defined stages in a research cycle: (1) Theory and research setup, (2) Collection, (3) Analysis, (4) Delivery. In defining skills we aimed at the middle level of abstraction: skills do not relate to particular implementations, platforms or software, but embody general practices and activities, while remaining useful and distinct (e.g. corpus building, classification, statistical modelling). To understand supply we have manually annotated current offers in a sample of European university courses in Digital Humanities and summer school workshops, tying them to the grid.

Review of the Data Landscape

Review report documenting the state of literary data (https://zenodo.org/record/6861022) was our most important achievement in this period. To enhance public outreach, the principal author explained the results of the study in layman’s terms: https://www.youtube.com/watch?v=jrJGTSWHuF0.
This review focuses on intellectual access, i.e. providing guidance for finding and sharing literary data, and consisting of collecting and analysing literary corpora, available formats, tools, and metadata. We will next create an exploratory catalogue of literary corpora and provide a transformation matrix/toolbox for solving

common issues. Our point of departure is the abundance of existing data, let alone their diversity or heterogeneity, as well as substantial differences in design and underlying concepts. Therefore, the document covers the definitions of text (is it a source, an edition, a data set? see chapter 3), the purpose of a corpus (e.g. general, reference, or monitoring corpora, special purpose corpora; see chapter 4), central considerations or criteria regarding the construction of a corpus (sampling, balancing, representativeness, annotation model(s), data format(s); see likewise chapter 4). How can one acquire data without transgressing ethical or legal boundaries (see chapter 5)? We ask: How can we assist literary scholars in searching for and finding existing data that are relevant to their own research questions? And additionally, what kind of research question is relevant concerning the present-day state of the data landscape and literariness and textuality?

Inventory of Existing Data Sources and Formats

Here we compile a comprehensive overview of the landscape of literary corpora and sources currently available, describing our methodological approach and analysing the various challenges encountered in the effort to collect information about these resources and consolidate them into a structured form. Based on an initial inventory of 86 corpora or corpus sets, we exemplify their wide variety with respect to structure, context and purpose, and consequently the differing modes of provisioning. We also propose a technological path towards making this information searchable via a central discovery catalogue by discussing principal design decisions regarding the data model and the technology stack needed for such a task.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984. Find out more about the team here: https://clsinfra.io/about/ourresearchers/.

Bibliography

Ciotti, Fabio (2014): „Digital literary and cultural studies: the state of the art and perspectives". Between 4/8, 1-17. https://doi.org/10.13125/2039-6597/1392

Eder, M. / **Rybicki, J.** / **Kestemont, M.** (2016): "Stylometry with R: a package for computational text analysis". R Journal, 8(1): 107-21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Fischer, Frank / **Ingo Börner** / **Matthias Göbel** / **Andrea Hechtl** / **Christopher Kittel** / **P. Miling** / **Peer Trilcke** (2019): "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama", in: Book of Abstracts of the Digital Humanities Conference 2019. Utrecht: ADHO.

Heiden, Serge (2010): "The TXM Platform: Building Open-Source Textual Analysis Software compatible with the TEI Encoding Scheme", in: 24th Pacific Asia Conference on Language, Information and Computation (pp. 10 p.). Sendai, Japon. Retrieved from http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24_sheiden.pdf

Mrugalski, Michał / **Odebrecht, Carolin** / **Charvat, Vera** / **Börner, Ingo,** / **Durco, Matej** (2022): CLS INFRA D5.1. Review of the Data Landscape. Zenodo. https://doi.or-g/10.5281/zenodo.6861022

Lisanne M. van Rossum / **Artjoms Šeļa** (2022): *CLS INFRA D4.1 Skills Gap Analysis* . Zenodo. https://doi.or-g/10.5281/zenodo.6421513.

Wilkinson, Mark D. / **Michel Dumontier** / **IJsbrand Jan Aalbersberg** / **Gabrielle Appleton** / **Myles Axton** / **Arie Baak** / **Niklas Blomberg** (2016): "The FAIR Guiding Principles for Scientific DataManagement and Stewardship", in: Scientific Data 3(1). https://doi.org/10.1038/sdata.2016.18.