Abstracts DHBenelux 2017 conference

Posters

1. Accelerating Linguistic Research through Advanced Digital Methods

Jan Odijk, Utrecht University

In this poster + demo I demonstrate two web applications that have been developed in the context of CLARIN-NL and CLARIAH for searching in treebanks (text enriched with syntactic structures). I aim to demonstrate that carrying out humanities research (linguistics, in this case) can be accelerated significantly by the use of such digital humanities applications. In particular, I will demonstrate the and GrETEL applications, including the recent extensions made by our team. I illustrate this by analyzing the Dutch words *heel*, *erg* and *zeer* ('very').

These words are (near-)synonyms but they show differences in their modification potential. This can be illustrated with the examples in (1). In (1a) the words *heel, erg* and *zeer* modify an adjectival predicate: all are correct. In (1b) these words modify a prepositional predicate which is synonymous with the adjective: *heel* cannot be used in this way (indicated by the *).

(1)

a. Zij is heel / erg /zeer zwanger 'she is very / very /very pregnant' (modification of adjective)

b. Zij is *heel / erg / zeer in verwachting `she is very / very /very in expectation' (modification of prepositional predicate)

It can also be illustrated using the word *verbaasd* `surprised', which can be an adjective or a form of the verb *verbazen* 'surprise'. When used as an adjective, *heel* can modify it, but when used as a verb *heel* cannot modify it: *Ik ben heel / erg / zeer verbaasd* v. *Dit heeft mij *heel / erg / zeer verbaasd*. The same restriction holds for the English translations of these examples when one uses the word *very* (*I was very surprised* v. **This has surprised me very*).

This raises many questions, in particular how children can acquire the relevant difference (Odijk 2011): they have to acquire this difference without recourse to semantics and they have to end up in a state in which they 'know' that *heel* I can**not** modify verbal or prepositional predicates.

In earlier work (Odijk 2015, 2016) I analyzed the modification potential of these words in a range of corpora, including the Dutch CHILDES corpora, crucially using the applications mentioned. Unfortunately, for some phenomena no or hardly any instances were found, which makes it difficult to draw any firm conclusions.

In this poster I will show how easy it is to obtain the data that are relevant for investigating this problem by the applications mentioned, and I compare the presented applications to similar applications such as LINDAT PML-TQ (Štěpánek & Pajas 2010), Tündra and INESS (Rosén et al , 2012).

I present results of an analysis of these words in an annotated collection of texts written for primary school children (Basilex: 13.5 million tokens, Tellings et al. 2014), to investigate whether we do find enough data in these larger corpora, and in the Wikipedia part of LASSY-LARGE (145 million tokens) to obtain an impression of what amounts of evidence we can expect at all with regard to these phenomena.

References

Jan Odijk. (2011). "User Scenario Search", internal CLARIN-NL document, April 13, 2011. [docx]

Jan Odijk (2015). 'Linguistic Research with PaQU' *Computational Linguistics in The Netherlands journal* 5, p. 3-14. [pdf] [url]

Jan Odijk(2016). `A Use case for Linguistic Research on Dutch with CLARIN', in K. De Smedt (ed.), 2016,*Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*, 45-61. Linköping University Electronic Press. [URL] [pdf]

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik (2012). An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco (eds.) *META-RESEARCH Workshop on Advanced Treebanking at LREC2012,* pages 22–29, Istanbul, Turkey, May 2012.

Štěpánek Jan, Pajas Petr (2010). Querying Diverse Treebanks in a Uniform Way, in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Copyright © European Language Resources Association (ELRA), Valletta, Malta, pp. 1828-1835.

Tellings et al. (2014). BasiLex: an 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal* 4 (2014) 191-208.

Van Noord et al. (2013). Large Scale Syntactic Annotation of Written Dutch: LASSY. In Spyns & Odijk (eds.) *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme.* Berlin/Heidelberg: Springer.

2. Trismegistos People

The extended version

Yanne Broux (KU Leuven)

Trismegistos [TM] (www.trismegistos.org) started in 2005 as a platform to facilitate access to information about papyrological and epigraphic sources (800 BC – AD 800) in all possible languages and scripts from Graeco-Roman Egypt.

TM also intensively deals with places [TM Geo] and people [TM People]. Combined with the availability of the full text in open access repositories, and the development of Named Entity Recognition [NER], TM has made significant progress in creating gazetteers of anthroponyms and toponyms that occur in the ancient sources. Starting out with Greek papyri, we tackled the full text of some 50,000 documents found in the Duke Databank of Documentary Papyri, and distilled over 350,000 personal names from the texts dated before AD 500.

In 2010 the idea grew to expand the geographical scope to include the ancient world in general. Our eventual goal is to transcend the compartmentalization caused by the limits of traditional heuristics and hermeneutics and to gain new insights into the ancient world by facilitating cross-cultural and cross-linguistic research.

Thanks to TM's role in EAGLE (www.eagle-network.eu), we were able to apply NER to almost 500,000 Latin inscriptions from the Roman Empire. Since the NER procedure developed for the Greek papyri was not geared toward the specifics of the Latin language and its naming system in particular, a custom procedure was set up, combining an onomastic gazetteer and a rule-based approach.

The onomastic gazetteer consists of all possible inflected forms of individual names (currently 872,879 forms). This was matched to the full text of the Latin inscriptions, from which some 400,000 clusters containing personal names were distilled. Each name in these clusters was tagged with a case, and connecting words were marked as such, so as not to split up the clusters (e.g. *Lucius Pescennius Cai filius Papiria Victor Severianus* is nominative – nominative – genitive – *filius* – *tribus* – nominative – nominative). The results of this first phase are currently being checked to filter out

'noise' (e.g. emperors' names used in dating formulae) or resolve ambiguous cases (e.g *Marci* can be gen. sing. or nom. pl.).

In a next phase, these sequences will be matched to a pre-defined rulebook to assign the names in the cluster to specific individuals (person, father, former master, ...). The Roman naming system was rather unique, as a person generally had a family name besides one or more individuating names. Other identifiers, such as a patronymic, (former) master, or voting tribe, could also be inserted. For example, the sequence ablative – ablative – genitive – *libertus* – ablative refers to a person with the *tria nomina* with his former master. Moreover, given the structure of a Roman name (praenomen – gentilicium – cognomen), an extra step is built in to make sure that each of these components is tagged correctly. Since common praenomina and gentilicia are marked as such in the gazetteer, this can also be automated. The human quality control of this phase will involve checking the interpretation of the composition of the cluster, as well as identifying individuals mentioned more than once within a single text.

3. To a lexicon of eighteenth-century pietistic language

M. Visscher-Houweling MA
PhD candidate Religious History Vrije Universiteit Amsterdam

Prof. dr. F.A. van Lieburg
Professor Religious History Vrije Universiteit Amsterdam

Listen Pieter, you should not laugh now, but, together with me, investigate which words are used by the mystics […] and whether they all deserve to be mocked. 1

Speaking is Abraham Blankaart, a fictional character in the epistolary novel Brieven van Abraham Blankaart, ('Letters of Abraham Blankaart') written by Elisabeth Wolff-Bekker and Agatha Deken in 1789. Blankaart's reason for speaking is a dictionary written by another personage in the novel, Pieter. This 'Woordenboek der fijnen' ('pietistic dictionary') consists of words and phrases typically used by pietistic Protestants. These Pietists were distinguishable from other Protestants by various characteristics, one of them being their use of language. Unfortunately, we can only guess which words and phrases were included in the dictionary of Pieter. Wolff and Deken have mentioned the dictionary in their novel, but did not incorporate it. However, it is possible to reconstruct a lexicon of eighteenth-century pietistic language with the aid of the various epistolary novels of Wolff and Deken and some eighteenth-century spectatorial magazines and theological books. In these works words and phrases are mentioned which were, according to the authors, typically used by 'fijnen'. In this study, conducted as a research pilot at the Meertens Instituut, a digital lexicon of eighteenth-century pietistic language is created. In the various works mentioned above, all digitally available, the words and combinations of words which are typically pietistic according to the authors of the works, will be tagged. The tagged words are then entered in a lexicon.

To test the value and usability of the lexicon the words and combinations of words from the lexicon will be compared with other eighteenth-century sources. A comparison with some theological books of ministers known as Pietists will make clear whether the lexicon is a veracious reflection of their use of language. It is possible that the lexicon will not so much reflect their language, but rather the language of pietistic lay people. To find that out the lexicon will also be compared with some egodocuments of pietistic lay people. Of course, it is also possible that the lexicon does not reflect

[1] E. Bekker and Agatha Deken, *De brieven van Abraham Blankaart* III (The letters of Abraham Blankaart III; 's Gravenhage 1789) 331. [Own translation]

either the language of pietistic ministers or the language of pietistic lay people. The authors of the epistolary novels, spectatorial magazines and theological books can have attributed words and phrases to pietists incorrectly, which will be a valuable result of the research as well. If the lexicon is a veracious reflection of the pietistic language it will also be compared to the Dutch Bible translation of 1637, the 'Statenvertaling', and the Dutch metrical psalms of 1773. This comparison will reveal to which extent the pietistic language is based on these works. The 'Statenvertaling' and the Dutch metrical psalms of 1773 are both available in Nederlab, a digital environment for the historical analysis of Dutch language. Lastly the lexicon will be compared with a lexicon of twentieth-century pietistic language2 to clarify what differences and similarities there are between eighteenth-century and twentieth-century pietistic language.

4. The Standardization Survival Kit (SSK)

For a wider use of standards within Arts and Humanities

Marie Puren, Inria

The H2020 project PARTHENOS (Pooling Activities Resources and Tools for Heritage E-research Networking, Optimization and Synergies) aims at strengthening the cohesion of European research in Arts and Humanities. The involved infrastructures have to address the new challenges posed by the increasing amount of digital contents and tools, and to support the emergence of a next generation of digitally aware scholars in the Arts and Humanities. Various reports and statements - like Riding the wave in 2010 - has thus acknowledged the growing importance to develop a data-centered strategy for the management of scientific data. In this context, standardization becomes a necessity for researchers in Arts and Humanities.

The poster will present the work carried out for a) the Standardization Survival Kit (SSK), an overlay platform dedicated to promote a wider use of standards within Arts and Humanities; and b) a specific awareness package articulated around the "Why standards?" leaflet to make scholars understand the essential role of standardized methods and content for the reusability of research results.

The SSK is a comprehensive interface aiming at providing documentation and resources concerning standards. It covers three types of activities related to the deployment and use of standards in the Arts and Humanities scholarship:

- Documenting existing standards by providing reference materials
- Fostering the adoption of standards
- Communicating to research communities

The SSK is designed as a comprehensive interface to guide scholars through all available resources, on the basis of reference scenarios identified since the beginning of the project. The design intends to provide a single entry point for novice or advanced scholars in the domain of digital methods, so that they can have quick access to the information needed for managing digital content, or applying the appropriate method in a scholarly context. Take, for example, a scholar, with few technical skills, who wishes to work on textual resources with digital tools. The SSK will offer her/him to explore reference scenarios, enabling her/him to easily discover new materials concerning standards (e.g. TEI P5 and its subsets), such as bibliographic references, tutorials, prototypical examples, transformation

[2] The lexicon of twentieth-century pietistic language can be found in: C. van de Ketterij, *De weg in woorden. Een systematische beschrijving van piëtistisch woordgebruik na 1900* (The way in words. A systematic description of pietistic word usage after 1900; Assen 1972). It has been digitized by me.

stylesheets. It will thus accompany her/him throughout her/his project, from the transcription of the primary documents to their publication online.

The poster will also present the "Why standards" leaflet - key element of the SSK's awareness package. The leaflet integrates a short text and a cartoon, and aims to be the point of entry to the SSK, but also to the PARTHENOS website and the helpdesk. This awareness-raising cartoon targets scholars with few technical skills: it is designed to communicate the necessity of standardization in the scientific world, in a way that will appeal to a wide audience and give a more modern and less off-putting image of standards.

---

5. The PARTHENOS Infrastructure

Sheena Bassett, PIN Scrl., Prato, Italy.

PARTHENOS aims at strengthening the cohesion of research in the broad sector of Linguistic Studies, Humanities, Cultural Heritage, History, Archaeology and related fields through a thematic cluster of European Research Infrastructures.

PARTHENOS will achieve this objective through:

- joint policies and solutions, including provisions for cross-discipline data use and re-use,
- the implementation of common AAA (authentication, authorization, access) and data curation policies, including long-term preservation;
- quality criteria and data approval/ certification;
- IPR management, also addressing sensitive data and privacy issues;
- foresight studies about innovative methods for the humanities;
- standardization and interoperability;
- common tools for data-oriented services such as resource discovery, search services, quality assessment of metadata, annotation of sources;
- communication activities, and joint training activities.

Built around two ERICs from the Humanities and Arts sector, DARIAH and CLARIN, along with ARIADNE, EHRI, CENDARI, CHARISMA and IPERION-CH and involving all the relevant Integrating activities projects, PARTHENOS will deliver guidelines, standards, methods, pooled services and tools to be used by its partners and all the research community.

Main results achieved so far

PARTHENOS started with defining the User Requirements using Use Cases contributed by the many partners and from an extensive literature search based on reports addressing researcher's needs in the Humanities and Arts. The first "User Requirements" report was completed in February 2016 with an updated internal version released in September 2016. All the subsequent work taking place relies on the User Requirements to develop their contributions.

Common policies and implementation strategies which includes data/metadata quality. The FAIR principles are used being used to make a structured overview of existing and needed policies for research data which will be implemented as an interactive "wizard".

An online Standards Survival Kit (SSK) is supported by a printed "Why standards" guide. Cartoon characters from a fictional planet (Digitus) are being used to promote the standards message as these are independent of subject domain or culture.

The technical work is focussed on developing the generic cloud structure and virtual research environment which will ultimately form the basis of the PARTHENOS tools and services. The

PARTHENOS Entities Model forms the underlying framework that researchers will interact with to achieve their tasks in PARTHENOS. The key components consist of a searchable registry of data sources along with the tools to enable researchers to extract, reformat and analyse, annotate their required information along with services for the deposit and preservation of their research work.

Skills, professional development and advancement - a first module which was introduced at the European Summer University in Digital Humanities in Leipzig (July 2016) aimed at informing less experienced researchers about the objectives and benefits of Research Infrastructures and how they can use them.

Finally, PARTHENOS is being widely disseminated through the website and Twitter, presentations at European and national events, press releases and booths and is establishing international liaisons and relationships with related organisations and other Research Infrastructures.

Please see www.parthenos –project.eu for further information on all the mentioned outputs and access to the training modules.

6. Bridging the Gap between Material and Immaterial Cultural Heritage in PARTHENOS VREs

Emiliano Degl'Innocenti [Corresponding Author]
Consiglio Nazionale delle Ricerche
Opera del Vocabolario Italiano
emiliano.deglinnocenti@ovi.cnr.it

Achille Felicetti	**Maurizio Sanesi**	**Roberta Giacomi**
PIN **University of Florence**	**Società Internazionale per lo Studio del Medioevo Latino**	**Società Internazionale per lo Studio del Medioevo Latino**
achille.felicetti@pin.unifi.it	**maurizio.sanesi@sismelfirenze.it**	**roberta.giacomi@sismelfirenze.it**

Overview

PARTHENOS3 will offer to DH and CH researchers access a number of digital resources and assets available in the digital ecosystem. To fully support current research practices and workflows, and to allow researchers developing new - possibly innovative - research trends, a strong digital research infrastructure has been designed and deployed around two main components: D-NET4 and D-4Science5. While D-NET is providing a set of services for the construction of Aggregative Data Infrastructures, including data mediation, data mapping, data storage and indexing, data curation and enrichment, and data provision, D4science is the component allowing the creation of Virtual Research Environments. VREs are web-based, collaborative, working environments supporting various kinds of scientific workflows through the integration of datasets and services available in the infrastructure into a user defined chain6.

At the core of the PARTHENOS infrastructure stands the Joint Resources Registry, a catalog of the available digital assets that could be exploited in a VRE7. A strong cross-disciplinary landscape is made available by the PARTHENOS resources coverage, encompassing the following disciplines: History, Archaeology, Art History, Epigraphy, Literature and Linguistics, Philology, Cultural Heritage, GLAMs as well as other disciplines in the SS domain[8].

---

[3] http://www. parthenos-project.eu (last visited 04.30.2017)

[4] http://www.d-net.research-infrastructures.eu (last visited 04.30.2017)

[5] https://www.d4science.org (last visited 04.30.2017)

[6] Cfr.: T. Blanke, L. Candela, M. Hedges, M. Priddy, & F. Simeoni *Deploying general-purpose virtual research environments for humanities research* Philosophical Transactions-Royal Society Of London Series A Mathematical Physical and Engineering Sciences, 368(1925), 3813-3828

[7] Cfr.: N. Aloia, L. Candela, F. Debole, L. Frosini, M. Lorenzini, P. Pagano *PARTHENOS Report on Design of the Joint Resource Registry* http://www.parthenos-project.eu/Download/Deliverables/D5.2_Report_on_design_Joint_Resource_Registry.pdf (last visited 04.30.2017)

[8] Cfr.: S. Drude, S. di Giorgio, P. Ronzino, P. Links, E. Degl'Innocenti, J. Oltersdorf, J. Stiller *PARTHENOS Report on User Requirements*. https://goo.gl/3IwI5J (last visited 04.30.2017)

Scope

The scope of this experiment is to prove that by re-using the components (i.e.: datasets, authority lists, services - both generic and specialized - and tools) that are accessible through the PARTHENOS Registy and integrating them in a custom VRE, users are able to elaborate complex research questions and explore cross-domain paths. The expected outcomes of this experiment is a demonstrator of such a VRE, that could possibly be further extended to other domains in the SSH landscape.

Methodology

For this experiment we started with a research scenario where a user is willing to retrieve information about a given CH artifact, regardless the discipline, research focus and pov assumed (i.e.: tangible vs tangible / descriptive vs textual vs instrumental vs contextual aspects) by the actual data providers. Out of more than 90 use cases[9] we focused on commonalities between historical/philological and archaeological datasets. Leveraging on the concept of *place* and exploiting the related information available in ARIADNE[10] and CENDARI[11] – two of the PARTHENOS participating e-infrastructures – we explored existing relations (on the CH side) with buildings, monuments and written accounts of archaeological research and connected them with information on people (i.e.: authors), primary sources (i.e.: manuscripts, charters etc.), origins and provenances (on the DH side of the PARTHENOS dataspace). The above entities and relations were tracked on time and space.

Outcomes

Based on the above methodological assumptions, using the D4Science platform and mapping the ARIADNE and CENDARI datasets to the CIDOC-CRM[12] common semantic framework[13] using the 3M mapping tool[14], we've elaborated a set of cross-domain research questions, leveraging on the notion of *place* as intersection between the CH (ARIADNE) and DH (CENDARI) dataspaces. Then we've developed a focused VRE implementing communication and analytics tools to track the above entities and relations over time and space, to demonstrate the technical feasibility and scientific value of the experiment. The VRE in its final iteration will implement data editing (i.e.: coverage check), browsing, and specialized services (i.e.: natural language processing, named entity recognition, etc.) and will allow results to be exported in a standard format (i.e.: XML/RDF), for storage or further reuse. The VRE will be accessible via the D4Science PARTHENOS Gateway[15]

---

[9] Cfr.: S. Drude, S. di Giorgio, P. Ronzino, P. Links, E. Degl'Innocenti, J. Oltersdorf, J. Stiller *PARTHENOS Report on User Requirements* cit.

[10] http://portal.ariadne-infrastructure.eu/ (last visited 04.30.2017)

[11] http://www.cendari.eu (last visited 04.30.2017)

[12] http://www.cidoc-crm.org/ (last visited 04.30.2017)

[13] Cfr.: G. Bruseker, M. Doerr, M. Theodoridou *PARTHENOS Report on the Common Semantic Framework* http://www.parthenos-project.eu/Download/Deliverables/D5.1_Common_Semantic_Framework_Appendices.pdf (last visited 04.30.2017)

[14] Already available in the PARTHENOS infrastructure: https://mapping-d-parthenos.d4science.org/3M/ListEntity?type=Mapping (last visited 04.30.2017)

[15] https://services.d4science.org/group/parthenos (last visited 04.30.2017)

7. Names and character diversity in Dutch children's literature

Gerrit Bloothooft[1], Lucas van der Deijl[1], Richard Thiel
[1]UiL-OTS, Utrecht University, The Netherlands
contact: g.bloothooft@uu.nl

Dutch parents name their children according to name preferences shared by others from their socio-economic, cultural, linguistic or ethnic group (Bloothooft & Onland 2011). As a result, popular names can be attributed to specific name groups – 14 clusters of names that co-occur above chance within families and represent socio-economic groups that approximately share the same level of education, income, ethnicity etc. (Bloothooft & Groot 2008). The study of these patterns in naming practices thus contributes to our understanding of cultural distinction and the social construction of group identities (e.g. Leys 1974; Desplanques 1986).

Representation and the construction of group identities have also been a main topic of interest among literary scholars. Recent research on character diversity in Dutch literature shows that the representation of social groups in the contemporary Dutch novel is relatively homogeneous compared to the composition of the present Dutch society (Van der Deijl et. al 2016). These findings raise new questions on the extent in which the experience, identity and preferences of authors structurally shape the fictional world. Such questions are especially relevant to children's literature, a genre that has been studied and criticized extensively for its depiction of the extra-literary world or the inclusiveness of its intended audience (Joosen & Van Lierop-Debrauwer 2014; Van Lierop-Debrauwer 2013).

Whereas literary names are usually studied as a stylistic phenomenon or as an aspect of a character's representation (Van Dalen-Oskam 2009, 2013), this project quantitatively examines character names in order to show whether the representation of social-cultural domains in children's literature at large compares to the distribution of these domains in the Dutch society. Because first names may signal social group membership, the distribution of names of characters in modern children's books has been compared to the distribution of contemporary first names of children in society. To this end, 2,907 first names, their character representation and the age group of intended readers of 538 Dutch children's books were selected (by the third author and made available at URL1.

This concerns an estimated 4.9% sample of all children's books published in the Netherlands between 2011-2016 of which the representativeness was controlled against properties of all books, as provided by the National Library of the Netherlands (KB). With the exception of rare or fantasy names (15% of all names), each child name was associated to one of 14 name groups. The resulting distribution of name groups was compared to the same type of distribution derived from the first names of all 3.1 million Dutch children who were between 2 and 18 years of age in 2016 (obtained from the Dutch Civil Registration and available in the Corpus of First Names in the Netherlands (Meertens KNAW, URL1)).

Results show a close correspondence between both name distributions, which indicates that children's literature fairly represents not only the names of children in society, but also its social-cultural and ethnic composition. These findings suggest a difference with adult literature, and puts the ongoing debates on literary diversity and representation in a new perspective (e.g. Rouw 2015; Amatmoekrim 2015; Loontjes 2016). On a methodological level, this project demonstrates how the comparison of textual data (names) with contextual information (socio-cultural features of name groups) enables a new perspective on both text and context. As such, this approach could be replicated for various other literary corpora, with challenging options for Named Entity Recognition (NER) and automated entity linking.

References

Amatmoekrim, K., ‘Een monoculturele uitwas. De ondraaglijke witheid van de Nederlandse letteren’. *De Groene Amsterdammer*, 20-08-2015.

Bloothooft, G. & D. Onland, ‘Socioeconomic Determinants of First Names’. *Names* 59 (2011) 1: 25-41.

Bloothooft, G. & L.F.M. Groot, ‘Name Clustering on the Basis of Parental Preferences’. *Names*, 56 (2008) 3: 111-163.

Dalen-Oskam, K.H. van. ‘Names in novels: an experiment in computational stylistics’. *LLC: The journal of digital scholarship in the Humanities* 28 (2013) 2 (June): 359-370.

Dalen-Oskam, K.H. van. ‘Professor Nummedal is niet alleen. Een analyse van de namen in Willem Frederik Hermans’ *Nooit meer slapen*’. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 125 (2009): 419-449.

Deijl, L.A. van der, S.A. Pieterse, M.L.M. Prinse, R.J.H. Smeets, ‘Mapping the Demographic Landscape of Characters in Recent Dutch Prose: A Quantitative Approach to Literary Representation’. *Journal of Dutch Literature* 7 (2016) 1: 20-42.

Desplanques, Guy. 1986. ‘Les enfants de Michel et Martine Dupont s’appellent Nicolas et Céline.’*Economie et statistique* 184 (1986) 1: 63–83.

Joosen, V. (ed.), van Lierop-Debrauwer, H. (ed.) & Ghesquière, R. (ed.) *Een land van waan en wijs: Geschiedenis van de Nederlandse jeugdliteratuur*. Amsterdam 2014.

Lierop-Debrauwer, H. van, ‘Twee stappen voorwaarts, een stap terug: De academische studie van de jeugdliteratuur in Nederland en Vlaanderen’. *Vooys* 31 (2013) 3/4: 8-18.

Leys, Odo. ‘Sociolinguistic Aspects of Namegiving Patterns’. *Onoma* 18 (1974): 448–55.

Loontjes, J., 'De literatuur is achtergebleven in het masculiene tijdperk'. *NRC Next*, 19-05-2016.

Rouw, E., 'Literatuur blijft te wit'. *NRC Handelsblad*, 16-05-2015.

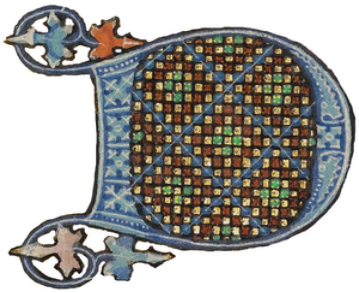
URL1: www.kjoek.nl

URL2: www.meertens.knaw.nl/nvb

8. Digital Medievalist

Presenting a Web-Based Community for Medievalists with Digital Media

Mike Kestemont (University of Antwerp)
Els De Paermentier (Ghent University)



Digital Medievalist (digitalmedievalist.wordpress.com) is an **international web-based community** for medievalists working with digital media. It was established in 2003 to help scholars meet the increasingly sophisticated demands faced by designers of contemporary digital projects. *Digital Medievalist* publishes an open access journal, sponsors conference sessions, runs an email discussion list, administrates active social media groups (e.g. *Facebook*) and encourages a best practice in digital medieval resource creation. With this poster, we would like to present the *DM* to a larger community in the Benelux Countries and inform the audience about our ongoing activities. Membership in *Digital Medievalist* is open to anyone with an interest in its subject matter, without regard to skill or previous experience in Digital Humanities or Medieval Studies. Participants range from novices contemplating their first project to many of the pioneers in our field. The entire *DM* community amounts to **over a 1,000 members worldwide**. The importance of the *DM* community is attested to in its status as the first Digital Humanities disciplinary-focus community of practice and a model for other groups (e.g. Digital Classicist, Digital Victorianist), and it serves a number of disciplinary fields, including digital humanities, medieval studies, cultural heritage, archaeology, literary studies, history, linguistics, and museum & archival studies.

A representative, yet non-exhaustive list of our activities includes:

- <dm-l@uleth.ca> is the ***Digital Medievalist*** **electronic mailing list**. Members use the list to ask for advice, discuss problems, and share information. The list's collegial atmosphere encourages a variety of conversations.
- *Digital Medievalist* (DM) is the project's **online, refereed Journal**. DM accepts work of original research and scholarship, notes on technological topics (markup and stylesheets, tools and software, etc.), commentary pieces discussing developments in the field, bibliographic and review articles, and project reports.
- Our website contains many useful sections, such as the **'Conference'** page, which offers a list of recent and upcoming conferences, colloquia, workshops and courses relevant to (digital) Medieval Studies.
- *DM* also has an active **Facebook group**, where news and discussion is posted relating to digital medieval issues.

Finally, we are extremely pleased to announce that *Digital Medievalist Journal* has recently joined ***The Open Library of Humanities,*** an academic-led, gold open-access publisher with no author-facing charges. With funding from the Andrew W. Mellon Foundation, the platform covers its costs by payments from an international library consortium, rather than any kind of author fee.

9. Data Huygens ING: Make the most of Humanities Data

Marnix van Berchum (Huygens ING)

Sebastiaan Derks (Huygens ING)

Ger Dijkstra (Huygens ING)

Jauco Noordzij (Huygens ING)

Lodewijk Petram (Huygens ING)

Digital Humanities has made great advances in recent years. But whereas digital techniques for working with relatively small, homogenous Arts and Humanities datasets are well-established, large collections of complex, heterogeneous data still provide a major challenge to the field. We present a new data infrastructure that is specifically designed to cope with the often complex and heterogeneous data inherent to many Arts and Humanities fields – especially those that follow a hermeneutic approach, such as History and Art History. Our infrastructure is ideally suited for managing, exploring, analysing, sharing, interlinking and enriching such datasets and will significantly contribute to the propagation of Digital Humanities techniques.

Central to the conception of our system are the notions of (1) 'provenance' and (2) 'multiple interpretations':

1. The infrastructure keeps meticulous track of data provenance. At dataset level, it documents general provenance information about e.g. data selection criteria and information extraction techniques. Furthermore, at data record level, it files the original source and all edits that have been made (e.g. by the dataset owner, or as result of a user-approved linkage with a record from another dataset). It thus creates a provenance trail that both gives insight into the origin of the data and allows to gauge its quality.
2. Arts and Humanities researchers should be able to disagree on every single data element, since data can be uncertain (e.g. due to contradictory sources) or contain subjective elements. Our infrastructure accommodates different views on a subject, and presents these side-by-side.

Our poster will be aimed at prospective users of the infrastructure (i.e. academic researchers in such fields as History and Art History) and we will therefore emphasise how they could potentially benefit from using the system. We will first provide a short explanation of the available functionality (uploading, managing, querying, analysing and downloading data). Furthermore, with a few appealing images related to a use case, we will convey that the infrastructure is very well suited for making connections between records from various datasets – either user-uploaded, contained within the system, or hosted externally (e.g. other Humanities Linked-Open-Data datasets or DBpedia). This will demonstrate how the infrastructure can be of great help to researchers in answering Who-, What-, Where- and When-questions: records with information on persons, concepts, locations and events can be enriched with data from a large selection of harmonised and standardised datasets, whilst the infrastructure allows for conflicting views on a certain entity to coexist and documents the provenance of all elements of a data record. Through our use case, we will also give insight into our data quality policy: to assure data interoperability, our Digital Data Management team offers guidance on issues of data standardisation and harmonisation. Moreover, close collaboration with our software developers guarantees that the data is seamlessly integrated in the system.

Finally, we will give an overview of the data already contained within the system and forthcoming additions from various national and international partners.

10. Artists in late medieval Ghent: a digital replication study

Yvonne Colijn, Marten Jan Bok, Harm Nijboer
University of Amsterdam

In recent years there has been an upsurge of interest in the replication of scientific research. A replication study might imply just redoing previous research to test its validity. More often it is done with the objective to test whether using a) the same methods on different or enriched data, or b) more advanced methods on the same data, or c) more advanced methods and different data leads to the same conclusions as the original study. As such, replication studies are key to the scientific method and the falsifiability of research results (Cf. Nosek & Errington 2017).

The increased interest in replication studies has not reached the humanities yet; at least, not massively. Still, there are good reasons to give replication a more prominent place on the research agenda. Past scandals have amply shown that replication in historical studies should not be taken lightly. Apart from the detection of fraud and dealing with milder forms of controversy, there is a growing need to reassess previous research to learn about the validity and value of digital research tools and digital data repositories for historical research.

In this paper we describe the digital replication of a prosopographical study by Els Cornelis on 248 artists in late medieval Ghent, published in 1987 and 1988. This study was largely carried out with data on paper, as was usual at that time. Fortunately, all data were included in a comprehensive appendix. We have entered these data in the ECARTICO prosopographical database at the University of Amsterdam.

In our paper, we will discuss:

- How digitizing data helped in reassessing the quality of the original data
- How it helps in enhancing the original data
- Digital methods and tools to reanalyze the data
- Whether the conclusions of the original study still hold after reanalyzing the data

We will include a short discussion on how digital replication studies such as ours can help to bridge the gap between digital and conventional research in historical studies.

References

Cornelis, Els (1987), 'De kunstenaar in het laat-middeleeuwse Gent I. Organisatie en kunstproduktie van de Sint-Lucasgilde in de 15de eeuw', *Handelingen der Maatschappij voor Geschiedenis en Oudheidkunde te Gent,* Nieuwe Reeks 41, pp. 97-128. <URL: http://ojs.ugent.be/hmgog/article/view/593/585>

Cornelis, Els (1988), 'De kunstenaar in het laat-middeleeuwse Gent. II. De sociaal-economische positie van de meesters van de Sint-Lucasgilde in de 15de eeuw', *Handelingen der Maatschappij voor Geschiedenis en Oudheidkunde te Gent,* Nieuwe Reeks 42, pp. 95-138. <URL: http://ojs.ugent.be/hmgog/article/view/259/251>

Nosek, Brian A. & Errington, Timothy M. (2017), 'Reproducibility in cancer biology: Making sense of replications,' eLife 6, e23383. <URL: http://dx.doi.org/10.7554/eLife.23383>

11. Dutch Overview Digital Humanities – Project Registry

Julia Noordegraaf, Claartje Rasterhoff, Ivan Kisjes, Romy Beck and Vincent Baptist
University of Amsterdam

DODH

The acronym DODH stands for Dutch Overview Digital Humanities, a digital resource that aims to map teaching and research activities related to Digital Humanities in the Netherlands.[16] DODH consists of two components. First, a **Course Registry** with an overview of Digital Humanities teaching activities in European countries. For Dutch speaking countries it covers courses in the Netherlands and Flanders. The Course Registry is part of an EC-wide DARIAH-resource.[17] And second, a **Project Registry** with an overview of accomplished and ongoing Digital Humanities Projects in the Netherlands or international projects in which a Dutch university or research institute has participated.[18]

The aim of DODH is to offer a chronological and thematic overview of the evolution of the field of Dutch Digital Humanities, of the involved institutions and persons, and of the disciplines that are represented. This can be relevant to researchers who study the evolution of their field and policy makers and funding programs who wish to assess the performance of DH initiatives that they have supported and funded. In particular, also smaller projects - executed at universities and research institutions have been indexed.

Execution

DODH started in 2014, and has been funded as a pilot project funded by CLARIAH (DARIAH_NL), and executed by the Erasmus Studio, Erasmus University Rotterdam (PI), in collaboration with University of Utrecht, and KNAW (DANS and eHumanities group).[19] DARIAH DE, the University of Cologne and the University of Göttingen have also supported the project. Recently, the research program Creative Amsterdam: An E-Humanities Perspective (CREATE) at the University of Amsterdam has taken over the Project Registry (not the Course Registry).[20]

The basis for the Project Registry was a combination of extracting seed information from the National Research Information System NARCIS, and crowdsourcing information among practitioners in DH. An attempt was made to validate the collected data by developing a review form and sending it to individual researchers that had been identified through the web as contact persons of the projects. The response however was so low, that a next round of very intensive web research and gathering information through personal contact had to be introduced. This has yielded sufficient information, up until recently it was not possible to tag each project according to the more refined Taxonomy of Digital Research Activities in the Humanities (TADIRAH).[21] In the last few months, we

---

[16] Stef Scagliola, Andrea Scharnhorst, Barbara Safradin, and Hendrik Schmeer (2016), *Dutch Overview Digital Humanities.* Demo presented at DHBenelux Belval, 9-10 June 2016, http://www.dhbenelux.org.; Stef Scagliola, Barbara Safradin, Almila Akdag, Hendrik Schmeer, Linda Reijnhoudt, Sally Wyatt, and Andrea Scharnhorst (2015), *Mapping Digital Humanities projects - A pilot of a DH project registry for The Netherlands*. Presentation given at the DH Benelux Antwerp June 8-9, 2015. https://www.slideshare.net/AndreaScharnhorst/a-pilot-of-a-dh-project-registry-for-thenetherlands/edit?src=slideview.

[17] https://www.clariah.nl/projecten/dodh/395-dodh#course-registry-2.

[18] http://dh-projectregistry.org.

[19] CLARIAH stands for Common Lab Research Infrastructure for the Arts and the Humanities and is a large Humanities infrastructure project in the Netherlands, funded by the National Research Agency NWO. http://www.clariah.nl/en.

[20] www.create.humanities.uva.nl.

[21] http://tadirah.dariah.eu/vocab/index.php

have simplified the data entry system to speed up information gathering and we aim to increase the number of TADIRAH tags associated with projects in the database.

Analysis

The Project Registry currently includes almost 400 projects, the first of which started in 1989. In last year's analysis of generated visualizations, a marked, but not unexpected, shift was observed, from hard core computational linguistics projects to initiatives in which the tools increasingly support opening up and analyzing data selected or created by humanities scholars and cultural heritage institutions [cf. note 1]. With this year's poster presentation we update the analysis of the data up to 2016, and evaluate the Project Registry's strengths and shortcomings. The DODH website currently offers basic functionality for exploration and descriptive analysis. It is, for instance, possible to query the data on institution, on project type, and on discipline. In this contribution we explore how such functionality may be expanded in the future, and how the data analysis can be enriched, for instance by means of international comparisons.

---

12. Golden Agents: Creative Industries and the Making of the Dutch Golden Age

Charles van den Heuvel, Harm Nijboer, et al.

The Golden Agents project will develop a sustainable research infrastructure to study relations and interactions between producers and consumers of 'creative goods' (works of art, books, writings, performances, etcetera) in the Dutch Golden Age. The project will link distributed, heterogeneous resources (both existing and new) so that researchers will be able to connect sources, agents, images, objects and texts from different resources in a new and meaningful way. For this purpose we will design a multi-agent platform that uses the various RDF schemes related to distributed datasets in combination with ontologies developed by domain experts. Golden Agents is a joint infrastructure project by Huygens-ING, University of Amsterdam, VU University Amsterdam, Utrecht University and the Amsterdam City Archives and is funded by the Netherlands Organisation for Scientific Research (NWO). Our poster will describe the envisioned architecture of this infrastructure and the key technologies (RDF, semantic reasoning, multi-agent technology) behind it.

---

13. Applications of the Diachronic Semantic Lexicon of Dutch (DiaMaNT) to historical document retrieval

Katrien Depuydt (katrien.depuydt@ivdnt.org) en Jesse de Does (jesse.dedoes@ivdnt.org)
Instituut voor de Nederlandse Taal

The Dutch language has been described extensively in the comprehensive historical dictionaries of Dutch [6]. The dictionaries provide the core material for the diachronic computational lexicon of Dutch (GiGaNT), that can be used to support search in historical texts by users without (expert) knowledge of historical spelling variation: when searching for *slager* ('butcher') the user also gets the morphological and spelling variants like *slaeger(s) slaegher(s)* or *slegher(s).* However, when a user studies the history of the butcher's trade, it is not immediately obvious from the way these traditional dictionaries are structured that one has also to look for *vleeschhouwer* or *beenhouwer* or *beenhakker.* And it is only after reading the complete articles that a user learns that *vleeschouwer* can also mean 'executioner', and *slager* 'a person who slays so.', be it though that in the case of *vleeschhouwer* the meaning *'executioner´* is derived from vleeschhouwer 'butcher',

while *slager* in contemporary meaning 'butcher' is derived from the meaning *'a person who slays someone'.*

The diachronic semantic lexicon DiaMaNT aims to enhance text accessibility and to foster research in the development of concepts, by interrelating attested word forms and semantic units (concepts), and tracing semantic developments through time. In the lexicon, the diachronic onomasiology, i.e. the change in naming of concepts and the diachronic semasiology, i.e. the change in meaning of words, will be recorded in a way suitable for use by humans and computers. The onomasiological part of the lexicon is meant to enhance recall in text retrieval by providing different verbal expressions of a concept or related concepts. The diachronic semasiological component (which charts semantic change), aims to enhance precision by enabling the user to take semantic change into account; the oldest meaning of *appel* is 'a fruit'.

A major challenge lies in coming up with a data resource and approach to use this resource, which will be beneficial to different types of research. To begin with, we looked into linguistic and digital humanities projects which either already use lexical resources and/or are focussing on the study of concepts and concept change in historical text. We also investigated the search behaviour of users of Delpher (www.delpher.nl). The prototype we will release in the course of 2017 will also contain a selection of vocabulary coming from these projects. We are currently also testing the lexicon in the context of two CLARIAH Pilot Research Projects, SERPENS (Contextual search and analysis of pest and nuisance species through time in the KB newspaper collection) and DB:CCC (Diamonds in Borneo: Commodities as Concepts in Context).

We see two potential applications that may benefit digital humanities:

1. Enhancing retrieval by using semantic relations in search
2. Word sense disambiguation in historical text

In this contribution, we share some first results of applications of the prototype lexicon in historical text retrieval, combining the lexicon data and distributional information obtained from historical corpora. Techniques (e.g. distributional approaches to language change [1], token-based vector spaces [2], word embeddings [3]) are evaluated using a specially developed benchmark dataset. The benchmark set consists of manually verified mentions of concepts in quotation material from the dictionaries and available historical corpora.

Bibliography

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop*, pp. 67-71.

Heylen, K., et al. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, **157**: 153-72.

Ignacio Iacobacci, Mohammad Taher Pilehvar and Roberto Navigli. Embeddings for Word Sense Disambiguation: An Evaluation Study. In Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016), Berlin, Germany, August 7-12, 2016.

Mooijaart, M. (2010). Doorlopende lexicografie: vier historische woordenboeken van het Nederlands in één databank'. *Neerlandica Wratislaviensia XIX*, 2010, 5-17.

14. A comparative analysis of the market for paintings in 16th and 17th century Antwerp and Amsterdam using ECARTICO

Harm Nijboer, Huygens ING

Marten Jan Bok, CREATE, University of Amsterdam

Anne-Rieke van Schaik, CREATE, University of Amsterdam

The ECARTICO database contains structured biographical data on more than 7,000 painters working in the Low Countries from the late Middle Ages to the early eighteenth century. All these data are structured in such a way that they can be queried for advanced geospatial an demographic analysis. Furthermore, because all individuals – and not just the painters – are represented as nodes in the data, the data allow also for various types of network analysis. The database is accessible through an on-line interface at http://www.vondel.humanities.uva.nl/ecartico/. Web users can browse through individual records, but they can also use several tools to visualize and analyze the data.

The ECARTICO data are not only derived from (art)historical literature, but also from original archival research. In recent years, extensive research in the Amsterdam baptism, marriage and burial registers has yielded a wealth of 'new' data on Amsterdam painters. And at the time of presentation ECARTICO will also have complete coverage of the records (*Liggeren*) of the Antwerp guild of Saint Luke. As a consequence, the coverage of the painter populations in these two cities is close to near complete.

With our poster, we will demonstrate the research possibilities of ECARTICO with a specific focus on a structural comparison between Antwerp and Amsterdam. The popular view of the development of these two artistic centers still holds that Antwerp flourished in the sixteenth century and was succeeded by Amsterdam in the Dutch Golden Age after the formers decay. However, analysis of the data tells a different story and it will also unveil some structural differences in the social coherence of the two painter populations that have not been highlighted in the historiography so far.

---

15. A Sea of Stone: Digitally Analyzing Jewish Funerary Inscriptions

Ortal-Paz Saar, Utrecht University, Department of History and Art History

This poster presents the first phase of a DH initiative titled PEACE: a Portal of Epigraphy, Archaeology, Conservation and Education on Jewish Funerary Culture. This phase focuses on epigraphic data: inscriptions recorded on stone, plaster, or gold-glass.

Funerary inscriptions shed light on an array of topics: life expectancy, personal names, family ties, official functions, and cultural norms of commemoration. They further illuminate matters such as inter-religious contacts and conceptions of the afterlife. In fact, from funerary inscriptions one can learn as much about the living as about the dead. This project employs digital means in order to allow the analysis of Jewish funerary inscriptions, and subsequent visualization of results. It relies on the following three databases, whose merging will allow the exploration of research questions on a chronological and geographical scale hitherto unattempted:

FIJI - Funerary Inscriptions of Jews from Italy (Utrecht University). Contains ca. 800 Jewish inscriptions, dated to the 2nd – 11th centuries CE, from Rome and Southern Italy.

Epidat (Steinheim Institute, Germany). Contains over 32,000 Jewish inscriptions, dated to the 11th – 21st centuries, primarily from Germany.

IIP - Inscriptions of Israel/Palestine (Brown University, USA). Contains ca. 3,000 inscriptions, dated to the 6th century BCE – 7th century CE, from present-day Israel and Palestine. The texts are mostly funerary, both Jewish and non-Jewish.

All three databases employ EpiDoc, a subset of TEI XML specifically designed for epigraphic documents. This fact facilitates the merging of the data, and permits the use of existing software and standard practices. Nonetheless, the analysis of these texts, often written in Hebrew or Aramaic, is no easy task. The RTL nature of these languages, the absence of vocalization, and the use of prefixes and suffixes, present numerous digital challenges, not only when performing complex text mining, but also when conducting simple searches.

The first phase of the project, still in progress, is conducted at Utrecht University, and involves the collaboration of historian Ortal-Paz Saar, the Utrecht University DH Lab, and Thomas Kollatz, the builder of Epidat. It consists in merging the data contained on the three databases, standardizing the search fields, and making the data uniform. This will allow to embark on the second phase, namely to search for terms and concepts (e.g. Eden, afterlife) and specific parameters (e.g. age, sex), consequently identifying historical patterns and tracing change over a long period of time. The research questions will pertain to four segments: Jewish commemorative norms, gender and sexuality, onomastics and emotions.

By applying to the epigraphic records computational methods of large data analysis, it will become possible to navigate this sea of stone in ways that researchers have not embarked on before.

Further reading:

Itai, Alon, and Shuly Wintner. 2007. "Language Resources for Hebrew." *Language Resources and Evaluation* 42:75–98.

Kollatz, Thomas. 2015. "epidat - Datenbank zur jüdischen Grabsteinepigraphik. Inventarisierung und Dokumentation historischer jüdischer Friedhöfe." In *Wenn das Erbe in die Wolken kommt. Digitalisierung und kulturelles Erbe*, edited by Eckhard Bolenz et al., 161–168. Essen: Klartext.

Reif, Stefan C., Andreas Lehnardt, and Avriel Bar-Levav (eds.) 2014. *Death in Jewish Life. Burial and Mourning Customs among Jews of Europe and Nearby Communities*. Berlin: De Gruyter.

Elliott, Tom, Gabriel Bodard, Hugh Cayless *et al.* 2006-2016. *EpiDoc: Epigraphic Documents in TEI XML.* Online material, available: <http://epidoc.sf.net>.

16. Trends in lexical diversity of the Troonrede, 1946-2016

Hugo Quené, Utrecht institute of Linguistics OTS, Utrecht University

The "Troonrede" is the annual speech by the head of state (Queen or King) of the Netherlands, spoken at the opening of the parliament session in September. This speech is generally written by the ministers of the national government, with editing by the prime minister and by the head of state. Because the context of the speech is ceremonial (and thus resistant to short-term trends in idiom and linguistic register), the contents of the subsequent annual speeches may reflect meaningful longitudinal trends in verbal expression. For two reasons, lexical diversity in the Troonrede is hypothesized to increase between 1946 and 2016: firstly because the Troonrede presents government plans regarding an increasingly complex and interconnected society, and secondly because the linguistic register of the Troonrede may have become more varied (also using less formal words and idioms) since the 1970s. This study investigates whether lexical diversity does indeed reflect such meaningful longitudinal trends, by means of quantitatively modeling the textual lexical diversity, using time (year) as a predictor.

A Measure of Textual Lexical Diversity (MTLD-MA, McCarthy & Jarvis, 2010, using factor 0.69 as assessed in pilot analyses) was determined for each Troonrede. Lexical diversity is typically expressed as the type-to-token ratio (TTR), which decreases as the text sample increases. Re-computing the TTR after each incremental word, the type-to-token ratio drops below 0.69 after about 119 words on average across all Troonredes 1946-2016, thus resulting in a grand mean MTLD-MA of 119. (Actual computations are more complex than suggested here.)

The resulting annual MTLD values of each Troonrede were fed into a statistical regression model with time (linear and quadratic) and speakers (4 since 1946, dummy coded with Juliana as baseline) as predictors. The optimal model explained 29% of the total variance in MTLD. This model shows no significant linear increase, contrary to predictions. However it revealed a significant quadratic component, indicating that lexical diversity decreased in the 1960s and 1970s and that it has increased since ca 1980. In addition to this trend, there was a significant increase in lexical diversity for Queen Beatrix.

These findings suggest that any increase in lexical diversity may have started (a) only in the late 1970s, and (b) more abruptly than hypothesized. These longitudinal trends may be explained in part by the wider variety of topics being addressed in the later editions of the Troonrede (e.g. domestic social relations, Europe, terrorism, refugees), rather than by a less formal linguistic register.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behaviour Research Methods*, 42(2), 381--392.

---

17. Andrea C. Bertino, Goettingen State and University Library

HIRMEOS - (High **I**ntegration of **R**esearch **M**onographs in the **E**uropean **O**pen **S**cience infrastructure**)**

Abstract

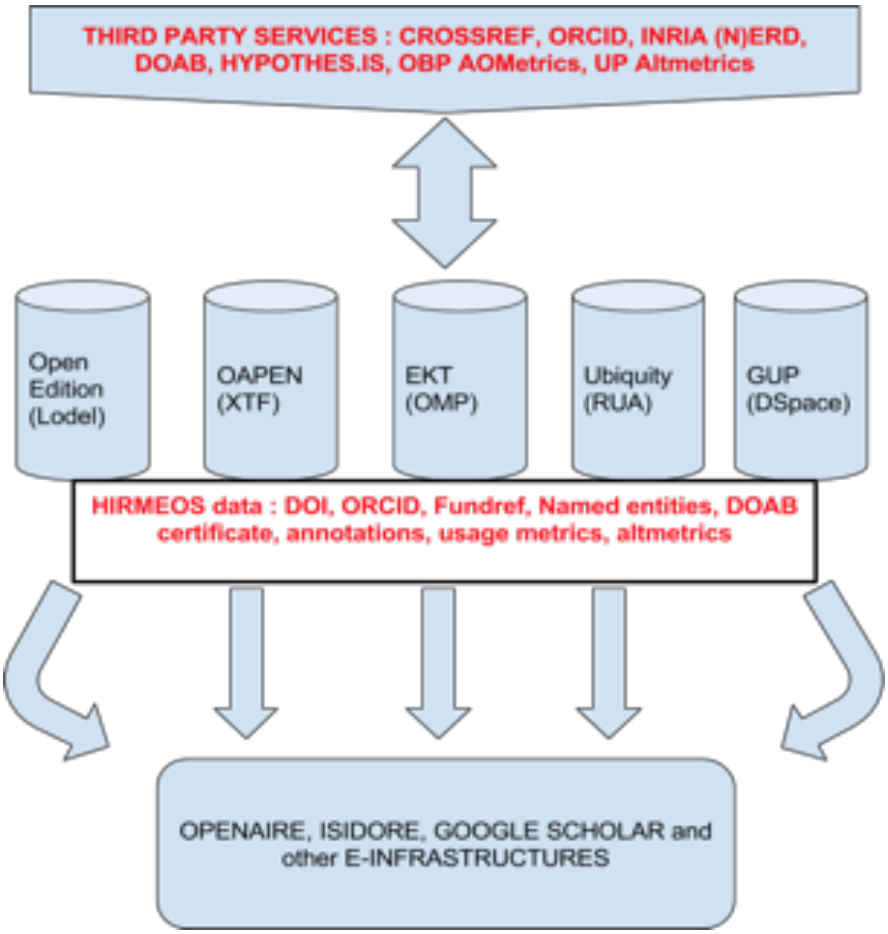
Open Science means more than mere Open Access, and Open Access does not mean just journals. Books are the way Social Sciences and Humanities communicate, but they still are peripheral in the Open Science environment. In our vision, developing Open Science goes beyond just releasing millions of Open Access documents in silos: it means creating bridges among countries and disciplines building an integrated trusted knowledge system.

High Integration of Research Monographs in the European Open Science infrastructure (HIRMEOS) is a 30-month project funded under the Horizon2020 Programme of the European Commission. It focuses on the monograph as a significant mode of scholarly communication in the Humanities and Social Sciences and tackles the main obstacles of the full integration of important platforms supporting open access monographs and their contents.

HIRMEOS and its related project, OPERAS-D, are part of a wider infrastructure project, OPERAS, aimed at integrating Social Science and Humanities within the European Open Science Cloud. In order to transform collections of passive documents into rich and interlinked content, HIRMEOS will work on innovative services on identification, entities recognition, annotation, and altmetrics, using when possible common standards.

HIRMEOS is based upon 5 Open Access books publishing platforms globally giving access to more than 7000 books. The platforms participating (OpenEdition Books, OAPEN Library, EKT Open Book

Press, Ubiquity Press and Göttingen University Press) will be enriched with tools that enable identification, authentication and interoperability (DOI, ORCID, Fundref), and tools that enrich information and entity extraction (INRIA (N)ERD), the ability to annotate monographs (Hypothes.is), and gather usage and alternative metric data. HIRMEOS will also enrich the technical capacities of the Directory of Open Access Books (DOAB), a most significant indexing service for open access monographs globally, to receive automated information for ingestion, while it will also develop a structured certification system to document monograph peer-review.



The partners of the project will develop shared minimum standards for their monograph publications, such that allow the full embedding of technologies and content in the European Science Cloud. Finally, the project will have a catalyst effect in including more disciplines into the Open Science paradigm, widening its boundaries towards the SSH.

The poster will give facts and figures about

- the involved platforms and their catalogues,
- the types of services to be implemented and the used standards,
- the expected benefit for the Social Sciences and Humanities players
- the global benefits for the Open Science environment

18. Maurizio Arfaioli (Medici Archive Project)

ITAF: Rewiring the Italian 'Nation' of the Army of Flanders (1567-1714)

Established in 1567 to crush a rebellion triggered by religious strife, and which through the years turned into a conflict of empires, the Spanish Army of Flanders rose to the challenge, becoming Europe's largest standing army since Roman times. A multi-lingual and multi-national army in which troops from all corners of the Spanish empire (and beyond) were called to serve.

By 1714 – when the Treaty of Utrecht ended the existence of the Spanish Netherlands – more than forty infantry regiments, dozens of companies of cavalry, several thousands of gentlemen adventurers, sailors, artillerymen, administrators, ecclesiastics and architects from all corners of Spanish Italy had 'passed to Flanders'. And that without mentioning the unaccounted-for women and children that followed them. Still, owing to the stigma that accompanied for a long time the memory of Spanish Italy, the Italian 'nation' in the Army of Flanders is nowadays largely forgotten, swallowed by an historiographical gap too wide to be filled with traditional – that is, non-DH – tools.

ITAF (Italian Troops of the Army of Flanders) is a document-oriented, NoSQL database (MongoDB platform) developed to 'rewire' the tangled web of military and social hierarchies which sustained the life of the Italian military 'nation' in the Low Countries through the effective integration of data extracted from a variety of archival and bibliographic sources disseminated between Italy, Belgium and Spain. To begin with, thanks to the use of a NoSQL database ITAF will allow to record and navigate the un-homogeneous mass of information (soldiers' variations in rank and wages, personal information etc.) contained in the volumes of the *Secrétairerie d'État et de Guerre* series in AGR Brussels, and to integrate it with data coming from other sources (personal correspondences, newsletters, wills, diaries etc.). This will permit the study of the life and professional trajectories of individual soldiers, as well as of entire groups – military, social, geographical etc.

ITAF's main technical features:

- MEAN Stack App (MongoDb, Express, AngularJS, NodeJs)
- News-centered Dynamic query masks for each type of news
- Non-homogeneous results display
- Source integration

The first trial version of this application/database is being developed in support a pilot study on the history of the unit we labeled as ITAF010, a *terzo* (the Italianization of the Spanish word *tercio* – an infantry regiment) that served the Spanish monarchy in the Low Countries from 1597 to 1689. The final version of this application is meant to be applied to the study of the entire Italian 'nation'.

ITAF is still being developed, but a first look at what it will (eventually) look like is available at this link: http://89.36.210.148:60000/

With a few adaptations (plus more time, resources, etc.), the use of ITAF could be easily extended to each of the other 'nations' (Spanish, Flemish-Walloon, British, German etc.) of the Army of Flanders or – more ambitiously – to the Army as a whole, offering a 'transversal' insight into the evolution of early modern European society and culture.

References

Glete, Jan (2002). War and the State in Early Modern Europe: Spain, the Dutch Republic and Sweden as Fiscal-Military States. London, Routledge

Gonzalez de Leon, Fernando (2009). The Road to Rocroi. Class, Culture and Command in the Spanish Army of Flanders, 1567-1659. Leiden, Brill

Hanlon, Gregory (1998). The Twilight of a Military Tradition: Italian Aristocrats and European Conflicts (1560-1800). London, UCL Press

Parker, G. (1972). The Army of Flanders and the Spanish Road 1567-1659. Cambridge, Cambridge University Press

19. DARIAH-BE: Sharing experiences and lessons learned: Top 5 tips from establishing a network of *DH Research Centres* in Belgium

Sally Chambers, Ghent Centre for Digital Humanities, Ghent University

Katrien Deroo, Faculty Library of Arts and Philosophy, Ghent University

Tom Gheldof, Leuven Centre for Digital Humanities, KU Leuven

Björn-Olav Dozo, Centre Informatique de Philosophie et Lettres, Université de Liège,

Mike Kestemont, Platform{DH}, University of Antwerp

Wout Dillen, Platform{DH}, University of Antwerp

Digital Humanities is thriving in Belgium. As a Founding Member of DARIAH-EU, the *Digital Research Infrastructure for the Arts and Humanities*, our aim is to offer a sustainable portfolio of services enabling digital scholarship in the arts and humanities. To realise this DARIAH partner institutions are encouraged to establish *Digital Humanities Research Centres* which together form a humanities-specific digital ecosystem, offering services both within their own institutions and to other institutions in Belgium and beyond. This poster presents four DH centres in Belgium: three existing centres; the *Centre Informatique de Philosophie et Lettres* (CIPL, Université de Liège), the University of Antwerp's Platform for Digital Humanities (platform{DH}, UA) and the *Ghent Centre for Digital Humanities* (GhentCDH, Ghent University) plus the *Leuven Centre for Digital Humanities* (LCDH, KU Leuven) which is currently being established. Finally, we share our experiences and lessons learned from establishing digital humanities centres in our own institutions and interconnecting them via the DARIAH network.

The *Centre Informatique de Philosophie et Lettres* (CIPL, Université de Liège), founded in 1983, has two main missions. Firstly to provide "technical support" for the Faculty of Philosophy and Letters and secondly, the development of databases and software for research purposes. Currently the centre is increasing its visibility and is strengthening connections with the Liège Game Lab (virtual reality) and the IFRES (Institut de Formation et de Recherche en Enseignement supérieur) regarding questions of teaching in a digital age.

The *Ghent Centre for Digital Humanities* (GhentCDH) is an interdisciplinary research centre facilitating digitally-enabled research in the Arts and Humanities at Ghent University and beyond. The GhentCDH offers advice and guidance throughout the research project lifecycle where digital tools, methods or collections are used with a specific focus on collaborative databases, digital text analysis and geospatial analysis. For the research data management and training we work closely with the Library Lab of the Faculty of Arts and Philosophy Library.

The University of Antwerp's Platform for Digital Humanities (platform{DH}), founded in 2010, aims to group together the University's Digital Humanities research by building a community of practitioners and by aggregating and disseminating individual projects that are conducted in the field. At the University, it is responsible for the Digital Humanities bachelor course, and for organising both a

yearly Spring (or Summer) Academy and a Lecture Series in DH. The FWO Scientific Research Community 'Digital Humanities Flanders' (DHu.F) is also coordinated by the platform.

The *Leuven Centre for Digital Humanities* (LCDH) will serve as a new hub for DH researchers at KU Leuven, alongside the current Task Force Digital Humanities. From 2017 onwards, the LCDH will encourage DH researchers to work more closely together, for example on collaborative projects applications. The LCDH will support the sharing of knowledge and digital tools and will work closely with the ARTES library co-organising DH workshops, lectures, summers schools and conferences as well as developing a long-term preservation strategy for research data from DH projects at KU Leuven.

In this poster, we will present our 'Top 5 tips' to consider when establishing a DH centre, drawing on our experiences in Liège, Ghent, Antwerp and Leuven. These will cover topics such as: the importance of a mixed team of experts, the high-visibility of the DH centre's (training) activities, fostering a DH community through joint activities (e.g., project proposals and publications), the strategic positioning of the DH Centre within the institution and the interconnecting of centres through initiatives such as DARIAH.

---

20. Analysing player decision-making of a moral dilemma through a computer vision analysis of Youtube gameplay videos

Stephanie de Smale, Humanities, Bram van den Brink, Remco Veltkamp, Johan Jeuring Computer Science - Utrecht University

We present our research-in-progress of analysing player decision-making through a quantitative automatic content analysis of gameplay videos of a war game scraped from Youtube. As Radde-Antweiler and Zeiler (2015) illustrate, a context analysis of gameplay content on video platforms should focus on three things: (1) the game; (2) the player's performance; (3) the comments on a video. For this research, we focused on the game itself. Although quantitative textual analysis (such as analysing youtube comments) is a more institutionalised practice, quantitative image analysis is not done frequently. First, using a Youtube scraper (Digital Methods Initiative 2016) that uses the Youtube API, we obtained a sample of 500 gameplay videos of a particular moral dilemma in the game *This War of Mine* (11 Bit Studios). Second, we use automatic image analysis to recognize objects (such as chosen characters or resources) and scenes in the frames of these videos to visualise player decision paths. The recognition is based on SIFT descriptors in the OpenCV computer vision software library. Thirdly, we are visualizing these decisions in behavioural trees, mapping the different routes taken by players.

Serious war games, such as this one, are seen as potentially valuable for their ability to promote cognitive and affective empathy (Darvasi 2016). Therefore, understanding what gameplay content is shared on video platforms and analysing what a player does in these games is potentially valuable for peace education and conflict resolution. In this research, we set out to analyse a moral dilemma of a sexually violent scenario, where the player may choose to intervene or not to intervene. Analysing this scenario, we are particularly interested in the choices leading up to this decision. Choices such as which playable character or which resource the player chooses creates different player paths, resulting in different play patterns.

From a humanities perspective, this analysis offers empirical research of non-researcher centric analysis of gameplay data. Within critical game analysis, most gameplay centric research tends to be a "close-reading" of gameplay, where the researcher adopts an auto-ethnographic perspective (e.g. Aarseth 2003; Lankoski & Björk 2015). This limit the study of games whose narratives highly depend on player decisions for developing its storyline. As others noted, playing research is tightly connected

to the way in which we understand games (Karppi & Sotamaa 2012). Our aim in this study is not to completely remove the researcher from studying the game. Rather, insights from playing research were used to distill key decision moments of this particular moral dilemma.

Augmenting this close reading, the study incorporates a digital humanities approach, also termed 'distant reading' (Moretti 2013). Through statistical analysis, distant reading 'aims to generate an abstract view from observing textual content to visualizing global features of a single or of multiple text(s)' (Jänicke et al. 2015). We build on Šisler (2016) who uses it to map game-rule systems. Our approach differs in two ways: firstly, by focusing gameplay instead of the formal system of the game we are able to analyse the game in action. Second, where distant reading statistically analyses texts, we analyse images. The output however, is similar, namely, we set out to develop scenario trees by conducting a micro-analysis of one gameplay event to classify how a player navigates this morally complex situation.

References

Aarseth, Espen. 2003. “Playing Research: Methodological Approaches to Game Analysis.” In *Proceedings of the Digital Arts and Culture Conference*, 28–29.

Darvasi, Paul. 2016. “Empathy, Perspective and Complicity: How Digital Games Can Support Peace Education and Conflict Resolution.” *United Nations Educational, Scientific and Cultural Organization | Mahatma Gandhi Institute of Education for Peace and Sustainable Development*.

Digital Methods Initiative. 2014. *Youtube Scraper*. Software.

Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.” *Proc. of EuroVis—STARs*, 83–103.

Karppi, Tero, and Olli Sotamaa. "Rethinking playing research: DJ HERO and methodological observations in the mix." *Simulation & Gaming* 43.3 (2012): 413-429.

Lankoski, Patri, and Staffan Björk. *Game research methods: An overview*. Lulu. com, 2015.

Moretti, Franco. 2013. *Distant Reading*. Verso.

Šisler, Vít. "Procedural religion: Methodological reflections on studying religion in video games." *new media & society* (2016): 1461444816649923.

Radde-Antweiler, Kerstin, M. Waltemathe, and X. Zeiler. "Video gaming, Let’s Plays, and religion: the relevance of researching Gamevironments." (2014): 1-36.

11 Bit Studios. 2014. *This War of Mine*. Software.

21. Hybrid query solution for humanities domain
Matteo Lorenzini

Introduction

During the last years, the production of digital datasets in the Cultural Heritage domain, has led to an exponential increase of information and, the structured repositories, have become the most used infrastructure for knowledge management and consultation towards different kind of system and platform ensuring a complete interoperability and reachability of data. In cultural heritage, thanks to the technology related to semantic web, we are able to manage and enrich our data using formalisms and data standards: digital libraries, digital archives and SPARQL-endpoint are some

examples. However, the fragmentation of data produced by different kinds of mapping methodologies and different representation of knowledge that needs to be managed, leads to some discrepancy during data retrieval between domains and results obtained.

A typical example is the study of an inscription: will be addressed by linguists in regards of the language; by philologists in regards of its text; by historians as a source; by archaeologists as material testimony of events and by conservationists as a piece of matter to be preserved and restored. In the light of that scenario, semantic interoperability and standardization are two fundamental elements able to guarantee the circulation of the knowledge inside a shared environment.

In the light of what I have described I'm going to present with this proposal the outline of my doctoral research which aims to set an extension to the SPARQL query syntax that allows to perform hybrid queries independently from metadata schema/format or infrastructure in order to:

- develop a query optimization techniques based on the use of description logic algorithms to match entities search clauses to appropriate repositories, combine retrieved results seamlessly, and reduce the query recall.
- improve the RDF graph obtained by SPARQL query with a semantic enrichment. The semantic enrichment will be based on CIDOC-CRM and will be useful to the distinction and identification of the same entities concepts in different domains i.e: an inscription considered as a source by the historians or as an archaeological evidence by the archaeologists. Methodology The development of the extension will be divided into two different parts: Design and implementation.

Design

The design phase will be characterized by the definition of the graph pattern and by the definition of the SPARQL query pattern. In the graph pattern will be defined the main entities usefull for the semantic model implementation and for the definition of the relations relation between the main entities. The graph pattern will be mapped on CIDOC-CRM and will represent the key for the further SPARQL query template. The SPARQL query pattern will be defined from the graph pattern as a set of triple patterns (subjetc, object, predicate) in OWL-DL syntax. The use of description logic level together with OWL will be usefull for the abstraction of the SPARQL query template composed of the logical operators able to infer and make reasoning process on the different nodes from the graph model. That allows to deduce the knowledge and information missing between two or more nodes of the graph. I.e Giovanni Piranesi=E21Person(architect Illustrator).

Implementation

The implementation phase will be mainly focused on:

Implementation of the ontology CIDOC-CRM based. Starting from the graph model the ontology will be extended using OWL-DL layer. Here will be formalized the axioms useful for matching the entities from different knowledge environments. Axiomatisation will be useful for:

- Automatic or semi-automatic integration of the resources between different repositories. SPARQL query consolidation considering also sub-graph from the main OWL graph model.
- Reasoning services in order to deduce the implicit knowledge. Development of the regular expressions based on tableaux algorithms. Regular expression and algorithms will be developed in order to guarantee the reasoning and inference useful for the semantic enrichment of the missing node from graph pattern. Development of the SPARQL query template. Sources are not limited to SPARQL endpoints, but also consist of documents containing RDF/XML, RDFa and other formats used to represent Linked Data. In order to satisfy the aims of the project will be combined two differents SPARQL query approaches: Partition-based and Federeted.

SPARQL query will be performed using apache Jena. The framework will provide both to the reasoning both to semantic enrichment of the resources discovered. During query compilation the federated SPARQL query is decomposed into a set of triple patterns (TP) TP = tp 1 , tp 2 , ..., tp i , where there are i triple patterns in the SPARQL query. Jena will provide to define an empity RDF graph populated by the data sources according to triple patterns mapping previously defined. Resurces will be finally semantically enriched according to the mapping schema.

Conclusion

Over the years in humanities domain, the semantic management of digital resource has became a common approach. Now the challange is how to make the knowledge till now defined available through all of those infrastructures. SPARQL end-point and registries are representing a great solutions: SPARQL end-point guarantee the integration of one dataset to the others according to Linked Data paradigm, registries aims to integrate the existing data in one common environment in order to improve the resource discovery according to one common mapping file. However, in both of cases, because of different kind of mappings approach and domain of interest, it's very difficult to check if the metadata are described and indexed uniformly according, for example, to their provenance.

My PHD project it is focused on the development of the hybrid solution previously described as a possible way to face the problems related to the resource discovery and integration. The expected results will be characterized by the possibility to:

- Improve the resource discovery: description logic expression should be able to guarantee a deeper level of description of the resources.
- Perform SPARQL query indipentently from metadata profile or data model.
- Infer the implicit knowledge from the discovered resources.
- Enrich semantically the resources discovered with the inferred knowledge.

References

Aloia,N., Papatheodorou, C., Gavrilis, D., Debole, F., and Meghini, C. Describing research data: A case study for archaeology. In Robert Meersman, Herv´e Panetto, Tharam S. Dillon, Michele Missikoff, Lin Liu, Oscar Pastor, Alfredo Cuzzocrea, and Timos K. Sellis, editors, OTM Conferences, volume 8841 of Lecture Notes in Computer Science, pages 768–775. Springer, 2014. Doerr, M., Theodoridou, M., Crmdig: A generic digital provenance model for scientific observation. In Peter Buneman and Juliana Freire, editors, TaPP. USENIX Association, 2011. Fernndez, M., Cantador, I., Lopez, V., Vallet, D., Castells, P., Motta, E., Semantically enhanced information retrieval: An ontology-based approach. J. Web Sem., 9(4):434–452, 2011 Loizou A., Angles, R., Groth, P., On the formulation of performant sparql queries. J. Web Sem., 31:1–26, 2015. Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., Manitsaris, A., A conceptual framework for metadata quality assessment. Universit¨atsverlag G¨ottingen, 104, 2008.

Pan, P., Beckmann, P., Havemann, S., Tzompanaki, K., Doerr, M., W. Fellner, D., A distributed object repository for cultural heritage. In Alessandro Artusi, Morwena Joly, Genevi`eve Lucet, Denis Pitzalis, and Alejandro Rib´es, editors, VAST, pages 105–114. Eurographics Association, 2010.

Peng P., Zou, L., Tamerzsu, M., Chen, L., Zhao, D., Processing sparql queries over distributed rdf graphs. VLDB J., 25(2):243–268, 2016.

Rademaker A., A Proof Theory for Description Logics. Springer Briefs in Computer Science. Springer, 2012.

Ruotsalo T., Hyvnen, E., A method for determining ontology-based semantic relevance. In Roland Wagner, Norman Revell, and Gnther Pernul, editors, DEXA, volume 4653 of Lecture Notes in Computer Science, pages 680–688. Springer, 2007.

22. Development of a IIIF-based digital corpus management and text analysis platform

Joke Daems, Sally Chambers, Christophe Verbruggen, Tecle Zere
Ghent Centre for Digital Humanities, Ghent University, Belgium

The digital text platform is part of the Flemish contribution to DARIAH Belgium (DARIAH = Digital Research Infrastructure for the Arts and Humanities). The goal is to create a platform for the collaborative management and discovery of digitised textual collections that allows digital humanities researchers to prepare their corpora (consisting of, for example, digitised newspapers and books) for textual analysis. The platform will enable researchers to browse and search the digitised collections compiled, cleaned, enriched and managed by the researchers themselves. Once the relevant research sub-corpus has been compiled, data export tools, using standardised open formats (such as XML, JSON, .csv, .txt, etc.) will enable researchers to export sub-corpus for analysis with existing digital text analysis tools such as MALLET, (http://mallet.cs.umass.edu/topics.php) for topic modelling, VOYANT (http://voyant-tools.org) for data visualisation or AntConC (http://www.laurenceanthony.net/software/antconc/) for concordance and textual analysis.

The platform has been conceived as part of a larger and modular virtual research environment service infrastructure (http://www.ghentcdh.ugent.be/projects/dariah-vl_vre.si). In a previous phase, possible frameworks and content management systems were tested, notably Islandora (a digital asset management system based on Fedora Commons and Drupal), but also Mediawiki and Omeka.

One of the main challenges of the envisaged new platform is the possibility to integrate a wider variety of possible textual data streams (including a scan workflow). In addition, user-friendliness, scalability, adherence to standards and facilitating the interoperability of data are key issues to be addressed. The platform will build on the existing IIIF format, the International Image Interoperability Framework. This format is used by some of the most important libraries and cultural heritage institutions in the world, therefore providing access to enormous collections of digital objects. As the name suggests, IIIF is mainly focused on displaying and annotating images. However, we fully endorse the IIIF-community's vision to develop an overarching interoperability framework for other data types, including all kinds of textual data. Benefits of the format include the interoperability, the ease of sharing images and annotations without the need to exchange files, and its support for multilingual data. In the months leading up to the conference, we will evaluate the existing IIIF-powered digital libraries and research projects and how they deal with practices of co-creation, data cleaning and enrichment of (structural) metadata. OCR improvement will become vital, as digital textual analysis can only be performed well on high-quality textual data. A related challenge will be combining the various input formats and converting them to different output formats required for analysis.

In our poster, we will present a summary of our experiences with and technical assessment of our previous Islandora installation, in addition to our survey of the existing corpus management solutions. As a way of conclusion, we will introduce the envisioned new version of the platform.

23. Timbuctoo: a linked open data infrastructure for complex, heterogeneous Humanities data

René van der Ark, Marnix van Berchum*, Bas Doppen, Ronald Dekker, Gertjan Filarski, Meindert Kroese, Martijn Maas, Valentina Maccatrozzo and Jauco Noordzij
All authors are affiliated with Huygens ING;
****corresponding author***

In this poster we present Timbuctoo, innovative software for managing, exploring, sharing, connecting and enriching Linked Open Data (LOD). This database system is specifically designed for academic research in the Arts and Humanities - research which often yields complex and heterogeneous data. It lives up to academic standards for working with such content: the software accommodates different views on a subject (i.e. interpretations of contradictory historical sources) and leaves the interpretation of the data to the researcher. Furthermore, Timbuctoo keeps meticulous track of data changes to facilitate provenance and version management.

The basic structure of the software is a set of REST API's on top of a linked data store (implemented on Neo4j), offering developers the opportunity to build clients interacting with these data. Currently the infrastructure provides:

- end user GUI's for uploading, searching and editing the data;
- a GUI for browsing the LOD web and connecting found entities to your own dataset;
- the ability to access the data as an RDF graph or as a REST (document oriented) datastore;
- an implementation of r2rml, enabling the conversion of existing relational data to RDF;
- various importers for binary formats;
- the ability to discover and download datasets from remote servers that contain ResourceSync descriptions;
- the ability to subscribe to changes on a dataset, which enables the creation of post-hoc data stores (e.g. MongoDB, MySQL) optimised for specific query patterns.

The software is fully Open Source; all code, documentation and issue tracking are available on Github: https://github.com/HuygensING/timbuctoo. Timbuctoo is being funded by Huygens ING and CLARIAH WP2; currently two instantiations of the software are running at http://data.huygens.knaw.nl and https://anansi.clariah.nl respectively. There are advanced plans for further implementations, creating a network of different instantiations of Timbuctoo which are able to use and synchronise each other data. Future collaborations with other national and international research projects include the *Cultures of knowledge* project of Oxford University.

24. Compiling CODECS: MediaWiki as a structured knowledge management platform for Celtic studies

Dennis Groenewegen (A. G. van Hamel Foundation for Celtic Studies; Wikibase Solutions)
dennis@vanhamel.nl

This presentation will offer a look at CODECS: Online Database and e-Resources for Celtic Studies (https://www.vanhamel.nl/codecs). CODECS is a web-based platform published by the A. G. van Hamel Foundation for Celtic Studies, a non-profit organisation based in the Netherlands. The main and most active part, which will be the subject of this presentation, is a research and teaching resource that marries the roles of bibliography, textual guide and catalogue of manuscripts for a broad range of primary textual materials available for the study of Celtic languages, literatures and cultures, currently with a strong focus on medieval and early modern sources associated with Ireland.

Its mission is to assist both scholars and enthusiasts in discovering and exploring these written sources and gaining a better appreciation of their historical contexts. Not only does the website offer 15,000+ entries for individual texts, manuscripts and publications: it also seeks to improve discoverability by mapping all relevant research data onto an easily extendable network of objects (entities and metadata) and object relations. Data currently being curated include agents, places, lexical items and literary motifs.

When requirements on the system to be used were formulated, three stood out most prominently:

1. flexibility of data modelling, storing and querying/visualisation
2. the means to facilitate editing / data input for a community of editors who are not necessarily technically minded
3. long-term sustainability of the software used and an active user base.

Our choice of package fell on MediaWiki as a robust, general-purpose framework in conjunction with a number of third-party extensions (plugins), notably Semantic MediaWiki (SMW) and Page Forms, with which to address (1) and (2). While then and now, the potential of this combination is relatively unknown to digital humanists, it has proved itself as an integrated, flexible and accessible solution for object- oriented and other structured approaches. The software has made considerable leaps in terms of both quality and community engagement, thereby satisfying our third requirement.

This presentation will single out key areas of development to demonstrate how the interplay between strategy and software has been critical to the ongoing growth of the project. It may be of interest, for instance, to anyone wishing to pursue a project involving granular classification and team effort; or anyone wishing to build a VRE (Virtual Research Environment) for specialist or interdisciplinary research where the data structure requires some pliability in order to grow with emerging insights.

In showing the internal maturity of the ecosystem as well as its open-ended nature (RDF, SPARQL, etc.), I also hope to sparkle the interests of those working on source-based technologies such as TEI XML and applications of image frameworks like IIIF. CODECS is ready for interaction.

Demos

1. Visualizing search behavior in digital libraries and archives using SWISH DataLab

Tessel Bogaard, Jan Wielemaker, Laura Hollink, Jacco van Ossenbruggen, Lynda Hardman
***Centrum Wiskunde & Informatica, Amsterdam, The Netherlands,* tessel.bogaard@cwi.nl**

Studying search behavior can help digital libraries and online archives to better respond to the needs and expectations of their users. It can help to evaluate search algorithms and user interfaces, or to identify potential gaps in the underlying document collection. Server logs store each individual user action in the online environment, but these low-level records can be hard to interpret due to a lack of context, and comparison between different types of behaviors is difficult. With SWISH DataLab[22] we group records into sessions (coherent sequences of actions by one user), and visualize them as graphs (Fig. 1).

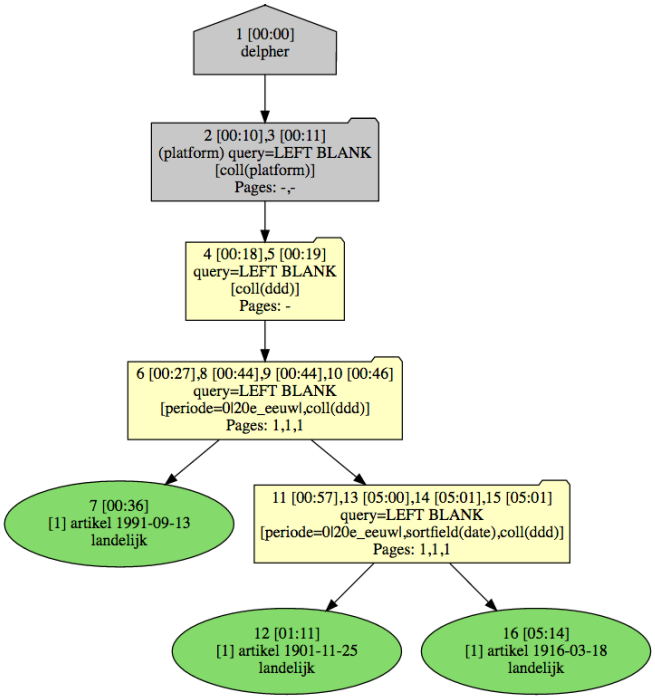


Fig. 1 Graph of a user visiting the home page (house-shape on top), followed by a search request (folder-shape), then adjusting this request a few times before clicking on results (ovals).

The nodes in the graph represent user actions, such as visits to the home page, search requests, or clicked results. The edges represent the transitions, ordered sequentially top to bottom, left to right. When a user revisits a page, or issues the same search request, this is represented as one node with multiple edges. Thus, the session graph gives a structural overview of what users do.

Collections from libraries and archives are usually described by professionally curated bibliographic metadata (such as publication date, origin or topic), often reflected in the search interface in so-called *facets*. Users can and often do apply facets from different metadata categories, adding an extra layer of information to their actions. We visualize the metadata values using colors, both for the search request (the facets), and for clicked results (using the bibliographic metadata). We can switch between metadata categories, for example a coloring by topic or by origin. This way we provide visual insight into metadata usage during search (Fig. 2).

[22] SWISH DataLab is based on SWI-Prolog (http://swish.swi-prolog.org/)

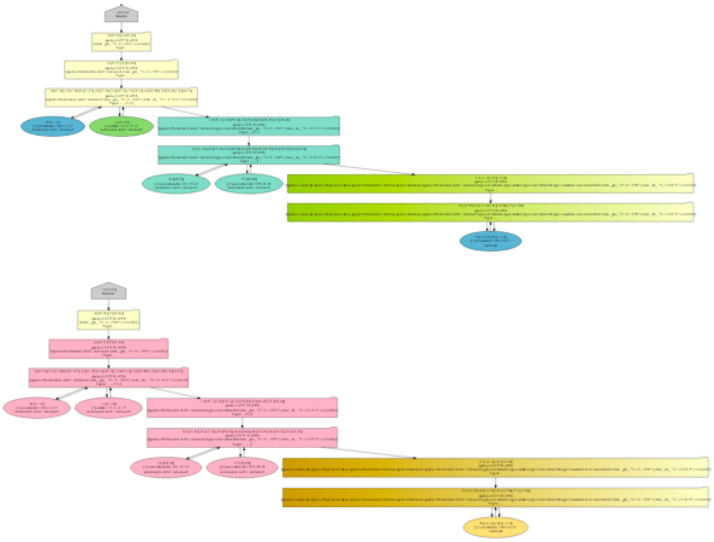


Fig. 2 The same session graph in two color modes (intentionally unreadable text in the nodes). The first shows the type of newspaper item (notice the three different colors, or item types, for the ovals, the clicks), the second for the distribution zones (two distribution zones in the clicks)

In this demo, we apply our technique to the log records of the National Library of the Netherlands (received under strictest confidentiality agreement). In these logs the IP addresses of the users are anonymized and only used to identify sessions; and the query is left out of the visualizations. This is a first step towards a more privacy-preserving method of data exploration, even so, the logs are privacy-sensitive and the demo will not be available online. We demonstrate how to explore and discover search behavior at a glance; and how to search for behavior satisfying certain criteria. Think of exploring sessions where certain facets have been selected, or a specific type of result was clicked.

The main aim of our work is to facilitate visual exploration of types of search behavior. As an added benefit, the quick visualizations support an iterative data cleaning process, and help defining abstractions over the data. Note that these abstractions can have a direct impact on further analysis: for example, not counting a search request revisited in the same session multiple times will help avoid overestimating the occurrence of these requests. Future work could include using graph properties (such as in- and out degree of nodes) as input for the analysis and prediction of search behavior.

Acknowledgement

The development of SWISH DataLab was partially supported by the VRE4EIC project that received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 676247

2. Corpus Upload and Metadata Analysis Extensions for GrETEL

Martijn van der Klis and Jan Odijk, UiL OTS, Utrecht University

Linguistic tools are primarily used only within the linguistic community, likely because they normally do not allow to import data, as well as lack options to deal with metadata. As part of CLARIAH WP3 we extended the treebank search engine GrETEL (Augustinus et al., 2012) with a possibility to upload corpora, as well as functionality to analyze and filter metadata.

The treebank search engine GrETEL allows researchers to use digital techniques (searching, parsing, and analyzing) on Dutch text corpora. GrETEL has a very user-friendly, example-based interface, but also allows queries in the XML query language XPath, and thus allows scholars on all levels to do digital humanities.

The corpus upload functionality we added allows users to upload an archived collection of plain-text files. The software will tokenize and parse these files using the Alpino dependency parser (Bouma et al., 2001), and import them into the XML database BaseX (Grün, 2010) for querying with GrETEL. Users can specify their corpus as private (only searchable for them) or publicly available. We are currently working on providing a wider range of input formats (e.g. CHAT, FoLiA, TEI).

For adding metadata to corpora, we use a format defined during development of PaQu, which allows users to add metadata in the running text (see http://zardoz.service.rug.nl:8067/info.html#cormeta for details). The software reads in the metadata and will create faceted search in GrETEL to allow users to both analyze and filter their search results. Users can change the facets to their liking, e.g. to use a range filter instead of checkboxes for numeric metadata.

After finding a result set of interest, this set can be further analyzed in an analysis interface. This interface allows to create pivot tables and graphs, which allows rapid insight into the data. The result set can also be exported to a spreadsheet format to allow further analysis in other tools.

We performed a case study on the CHILDES corpus (MacWhinney, 2000) to investigate the added value of this extension. This corpus contains transcribed text from parent-child interactions. We uploaded a small subcorpus (van Kampen, 2009) to our extension including, most importantly, metadata on the age of the speaker. This allowed us to search for appearance of certain syntactic constructions in first language acquisition. In particular, we found a sharp rise in the relative frequency of infinitive complements ("de pop moet slapen" – "the doll needs to sleep") at the age of four. The relative frequency stayed stable afterwards. Such rises in use could well signal this construction has been fully mastered.

The extension to GrETEL is created using CodeIgniter (a PHP web framework) and will be available as open-source software (MIT license) via GitHub: https://github.com/UiL-OTS-labs/gretel-upload

Keywords: treebank search engine – corpus upload – metadata – faceted search

References

Augustinus, L., Vandeghinste, V., & Van Eynde, F. (2012). Example-Based Treebank Querying. In *LREC* (pp. 3161-3167).

Bouma, G., Van Noord, G., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1), 45-59.

Grün, C. (2010). *Storing and querying large XML instances* (Doctoral dissertation).

van Kampen, J. (2009). The non-biological evolution of grammar: Wh-question formation in Germanic. *Biolinguistics*, *3*(2-3), 154-185.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

3. AnnCor: A treebank of spoken child Dutch

Pieter Hussaarts, Meie Otten, Remco van der Veen, Marjo van Koppen, Jan Odijk
Utrecht University

Language corpora are an almost inexhaustible source of language data for researchers. Corpora are frequently used by researchers to obtain, for example, information about word frequency and probabilities of words occurring in certain environments. Syntactic treebanks can be used by syntacticians to obtain information about sentence structures, and child speech corpora can even be used in language acquisition studies. However, as of yet, no syntactic treebank of Dutch child speech exists. AnnCor aims to add precisely this to the existing body of corpora.

More precisely, AnnCor is a treebank project of Utrecht University which uses the dependency parser Alpino (Van der Beek et al., 2002) for initial parsing of the Dutch corpora within CHILDES (MacWhinney, 2000). This online Alpino parser (together with the TrEd application) is a digital tool that can be used for research in the field of humanities on data that is stored digitally. Alpino was developed for written Dutch and can correctly parse 90% of written adult Dutch (Renckens, 2011). The Alpino parser is a perfect example of a digital tool that makes it possible to investigate research questions from the humanities with large quantities of digital language data.

However, since Alpino was developed for written and adult language, it is less suitable for SPOKEN (and) CHILD language. Within the AnnCor project, we are currently in the process of parsing the Van Kampen corpus. This corpus within CHILDES is a longitudinal learner corpus that has recorded the utterances of two children. As spoken language is relatively imperfect when compared to written language, and as child speech contains utterances that are structurally ungrammatical or at least deviate from the adult standard, many utterances in the corpus invite errors from the Alpino parser, leading to unsatisfactory analyses. In order to achieve a very high percentage of correct analyses, the potentially problematic utterances have to be checked individually and manually, and necessary alterations have to be made. This process involves multiple checks from different people and a check programme (ACE: AnnCor Check Engine) that is employed on a portion of the utterances to highlight possible violations and implausible structures. However, the process is still challenging and difficult decisions sometimes have to be made in order for the digital tool to work properly to generate syntactic trees for spoken (child) language. The decisions and alterations are therefore always documented, so researchers using the searchable treebank of child speech can judge our interpretations for themselves.

Upon completion, AnnCor will be a fully searchable treebank of child speech. It will increase the accessibility of CHILDES for researchers by offering thousands of fully annotated and parsed child utterances, and thus, by using digital tools, the project contributes to better digital data and therefore better digital humanities. Because of the longitudinal nature of, for instance, the Van Kampen corpus, language acquisition researchers will be able to investigate the syntactic development of children in a way that was not possible before.

References:

van der Beek, L., Bouma, G., Malouf, R., van Noord, G. (2002). The Alpino dependency treebank. *Language and Computers* 45(1), 8-22.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Renckens, E. (2011). Mens en computer ontleden even goed. Retrieved from https://www.nemokennislink.nl/publicaties/mens-en-computer-ontleden-even-goed

4. If buildings could talk

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

Menno Den Engelse, Islands of Meaning, menno@islandsofmeaning.nl

Richard Zijdeman, International Institute for Social History, richard.zijdeman@iisg.nl

Purpose

Buildings are an important part of cultural heritage, but we know surprisingly little about them. From archival records, we often know what architect designed the building, its construction date and sometimes its 'end date' (i.e. when it was demolished). All of these concepts can be requested as Linked Open Data too, for example through DBpedia [1]. However, we know little about people who lived, worked or relaxed in these buildings. By integrating knowledge about buildings and their inhabitants from different sources, we can facilitate more complex historical research on these buildings. Our demo presents a prototype application that identifies buildings and integrates information about these buildings from various sources.

Issues

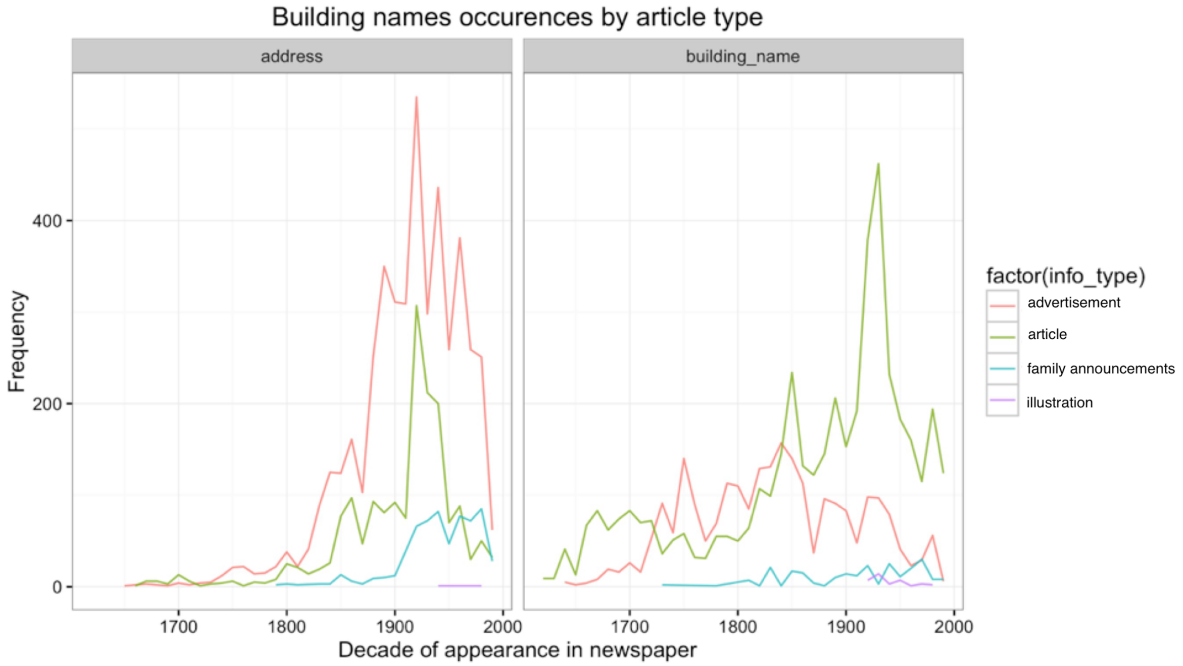


Figure 1. Variation in use of building name and address by newspaper article type

We identify two issues that make augmenting information on historical buildings difficult. The first issue concerns the identification of buildings sometimes by name sometimes by address. For example, both "Rijksmuseum" and "Museumplein 1, Amsterdam" can refer to the building with geographical coordinates 52.3600° N, 4.8852° E. Figure 1 shows the number of referrals to building names and addresses over time as taken from Delpher.nl: both ways of referring to buildings were

common in newspapers over the past two centuries. Furthermore, building names are not always unique; Amsterdam has for example known at least five structures named 'Haarlemmerpoort'.23

The second issue concerns extracting information related to buildings from textual sources. Figure 2 shows a sharp increase in the occurrences in Delpher of an address in Amsterdam. Content analysis reveals the reason behind the remarkable change. At first, the advertisements aimed to hire (kitchen) maids, whereas later on the advertisements were on title deeds, indicating a real estate agent had moved in (and got married somewhat later).

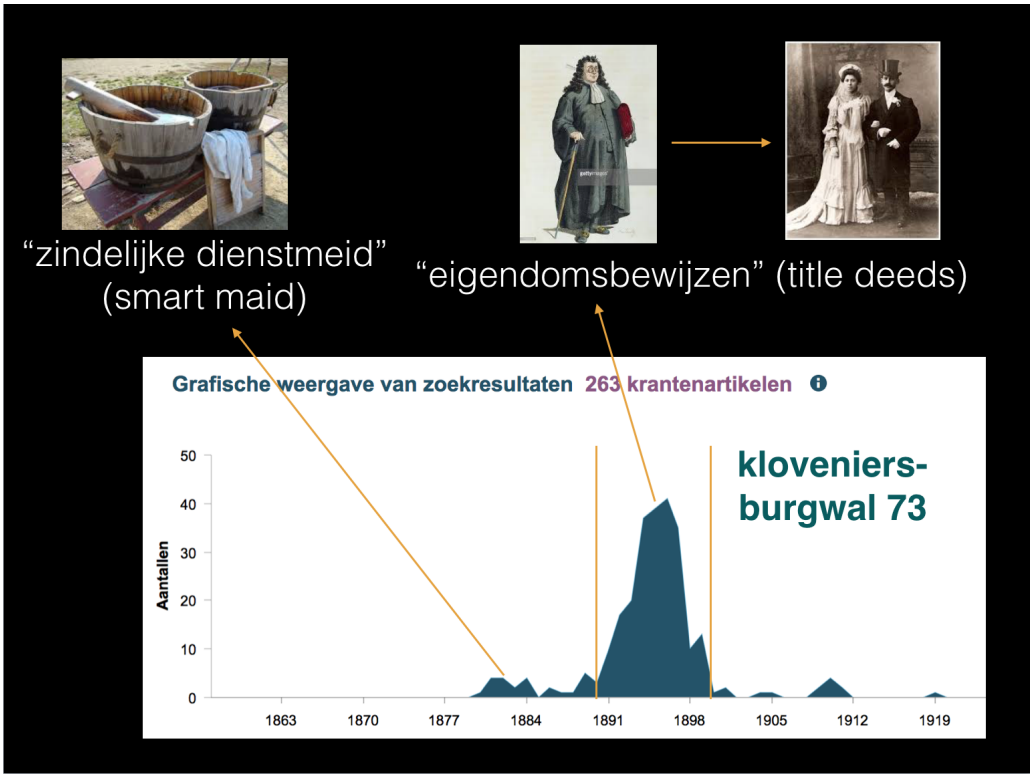


Figure 2. Occurrence and contents of advertisements related to an Amsterdam address.

Approach

Focusing on the city of Amsterdam for which many sources have already been digitised, we will create a dataset of names, linked to buildings and their geometries. Among the sources are the collections of the Amsterdam City Archives [2] and the Verdwenen Gebouwen dataset [3]. We will explore the possibility to add names extracted from newspapers. The dataset will be stored under an open license at the International Institute of Social History, and is available to other researchers and cultural heritage institutions. This allows for example collection owners to visualise collection items on maps, and, on an aggregated level, to connect items from different collections.

To identify buildings, we use the Basisadministratie Adressen en Gebouwen (BAG) Linked Open Data[4], providing a building id, address and building year. BAG applies to contemporary buildings,

[23] Tussen Haarlemmerpoort en Halfweg, Historische atlas van de Brettenzone in Amsterdam. Auteurs: Jaap Evert Abrahamse, Menne Kosian en Erik Schmitz. Uitgeverij Thoth, Bussum, oktober 2010. ISBN 978-906868-515-2

therefore information about historical uses of the building as well as for buildings that no longer exist will be added from other sources.

In the context of CLARIAH, we explore linking the enriched buildings dataset to information extracted from newspapers, aiming to build towards a rich and varied source on the history of buildings.

[1] http://dbpedia.org

[2] https://www.amsterdam.nl/stadsarchief/

[3] http://verdwenengebouwen.nl/

[4] https://www.kadaster.nl/bag

5. Demonstration ELMCIP Knowledge Base

Hannah Ackermans, research assistant at Bergen University, Norway

The ELMCIP Knowledge Base (KB) is a database of documentation of electronic literature - and critical writings, event, etc. related to electronic literature - that highly depends on the e-lit community. Everyone who has works to add to the database can request a contributor's account. This makes the KB a crowd-sourced database, which can be accessed by anyone interested in the topic. This helps everyone in the community in finding electronic literature they may not know yet and the database also provides the opportunity for quantitative research on the field of electronic literature by using all the metadata in the database for statistical analyses (i.e. Rettberg). The KB can thus be characterized as Digital Humanities 2.0 (Pressner), in which community and connection are key. Especially open-ended elements are taken into statistical research, such as tag words, in which contributors can fill in any word to describe entries. Consequently, qualitative and quantitative elements are intertwined in the documentation research: the structure of database favors subjectivity, ambiguity and contingency above elimination of "coder bias" – the influence of human researcher on results –, positing humanities values as a methodological strength.

Depending on the audience, this demonstration can be turned into a short workshop in which people get the opportunity to work with the KB themselves.

I can provide contributor accounts for all participants. After a short introduction to ELMCIP and the aims and practices of the Knowledge Base, I will explain the basics of adding personal record to the Knowledge Base. I will let participants add or edit a personal record for themselves and provide guidance while they do this. When they have done this, I will show how to add a creative work and critical writing. I will have a list of e-lit works that they can add to the KB in case they are not familiar with the field of electronic literature themselves. Adding and editing works in the KB can be learned easily as the database is very user-friendly. A short workshop can provide a good guidance to learn how the Knowledge Base works and how they can use it themselves in future research.

I will also give short demonstrations of the use of metadata in the KB into statistical analyses, showing some figures of representation of gender, nationality, tag words etc. in the knowledge base. I will also give people the opportunity of use the KB themselves to look up the metadata in which they are interested and do some short analyses.

This demonstration/workshop might be primarily interesting for people who have an affiliation with electronic literature or media art. However, on a more abstract level the use of the database and analysis in such an open community-based manner also contains knowledge and ideas that can be extrapolated to other fields of interest.

6. A Tool for Flexible and Transparent Text Processing Pipelines

Janneke M. van der Zwaan, Wouter Smink, Anneke Sools, Gerben Westerhof, Bernard Veldkamp, and Sytske Wiegersma

Many Digital Humanities (DH) research projects apply text mining tasks, such as sentiment analysis, named entity recognition, or topic modeling. These tasks all require the use of software. Fortunately, a lot of software is alreadyavailable, for example, NLTK [2], gensim [5]. However, using such software in actual research projects can be challenging. This challenge increases when researchers want to combine tools from different packages. The resulting scripts generally duplicate at least some text processing tasks (e.g., tokenization), and need to be adapted when used for new datasets or in different software or hardware environments. This affects research reproducibility and reuse of existing software. Using a standard for data analysis pipelines, such as Common Work-flow Language (CWL) [1], can help to solve these problems.

We introduce nlppln (NLP pipeline)[24], an open source software package that helps researchers create NLP pipelines using CWL. The advantage of using CWL is that any existing NLP tool can be integrated into a workflow, as long as it can be run as a command line tool. Existing frameworks for NLP pipelines are usually restricted to tools written in a specific programming language (e.g.,

GATE [3], or DKPro Core [4]). This is especially a problem for researchers working with non-English text, because tools for those languages may be only available in a different programming language. We believe that the flexibility and transparency introduced by CWL and the software really add something to existing text processing tools and frameworks for making NLP pipelines.

In order to be able to run tools and sequences of tools (i.e., pipelines), CWL needs a specification of a command line tool (i.e., a text processing step). nlppln helps creating those specifications for new and existing command line tools. The software also provides steps for (generic) NLP functionality, such as tokenization, lemmatization, and part of speech tagging. Third, the software provides functionality to convert existing NLP tools written in Python to command line tools (with associated CWL steps), and simplifies creating new steps, for example, for project specific tasks. Finally, the software helps users to combine (existing and new) processing steps into pipelines.

The demo features an example pipeline that removes named entities from text files. In addition, researchers are invited to create pipelines for their own text mining tasks.

References

1. P Amstutz, M R Crusoe, N Tijanić, B Chapman, J Chilton, M Heuer, A Kartashov, D Leehr, H Ménager, M Nedeljkovich, M Scales, S Soiland-Reyes, and L Stojanovic. *Common Workflow Language, v1.0*, 2016.
2. S Bird, E Loper, and E Klein. *Natural Language Processing with Python.* O'Reilly Media Inc., 2009.
3. H Cunningham, D Maynard, K Bontcheva, and V Tablan. GATE: an architecture for development of robust HLT applications. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 168–175, 2002.
4. R de Castilho and I Gurevych. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, 2014.
5. [5] R Rehurek and P Sojka. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, 2010.

24 https://github.com/WhatWorksWhenForWhom/nlppln

7. Distilling careers: augmenting biographies with occupational information

Antske Fokkens, Richard Zijdeman and Auke Rijpma

By the end of the 20th century, post-war trends of declining inequality halted or even reversed in many countries. The public concern over this issue is apparent in media attention and academic work on trends of economic and social inequality (e.g. Piketty and Ganser 2014, Clark 2014, Van Zanden et al. 2014). Current historical studies of career mobility often focus on linkage of structured data (tables) such as baptism records. More qualitative sources, such as biographies contain vital information as well, but are labour intensive to process. We propose a combination of Robust Semantic Parsing and Linked Data conversion tools to automatically derive career patterns from 35,000 biographies in the Biography Portal in the period 1815-1940.

This proposal meets the public concern and academic interest by evaluating whether Robust Semantic Parsing tools can be used to answer historically substantive questions on social inequality, namely:

1. How did career patterns in the Netherlands in the long nineteenth century look and how did they change over time?
2. To what extent do job advertisements in Dutch news papers reflect the hypothesized change from ascription to achievement for occupational attainment?

Both questions have recently been studied (e.g. Schulz & Maas 2010, Schulz, Maas & Van Leeuwen, 2015). Our aim is to 'replicate' those studies using innovative methods. Previous work focuses on linking events (birth, marriage and death records or census data) or base conclusions on small samples of biographies. Linkage approaches are limited because many variables introduce linkage biases. This does not apply to biographies, but their manual analysis is time consuming. We circumvent this through Robust Semantic Parsing Tools in combination with Linked Data converters to extract career paths from approximately 35,000 biographies from the Biography Portal. For this purpose we will apply **Dutch Robust Semantic Parsing Tools** for text mining, a set of open source tools provided through CLARIAH. Notably, we use:

- Word Sense Disambiguation (WSD)
- The simple tagger with occupation linking (HISCO tagger)
- NLP2RDF crystallization (the BiographyNet variation) The HISCO tagger identifies occupations mentioned in text and links them to their HISCO code. WSD resolv esambiguities(e.g . *broeder*canmean``monk''or``brother'').TheBiographyNetNLP2RDF crystallization links occupations to people. Once the occupations are retrieved and put into chronological order, we can add HISCAM scores to the occupations indicating whether occupations have a high or low standing in the occupational structure. By doing so we can create career patterns for the individuals in the Biography Portal and potentially feed them to back to the portal as life course visualizations. In addition to the substance information this will add to the Biography Portal, this demonstration will highlight the potential of augmenting structured (tabular) data with textual sources. Figure 1 illustrates the actual output of the semantic processing tools when applied to Kutusomo Ine's biography.

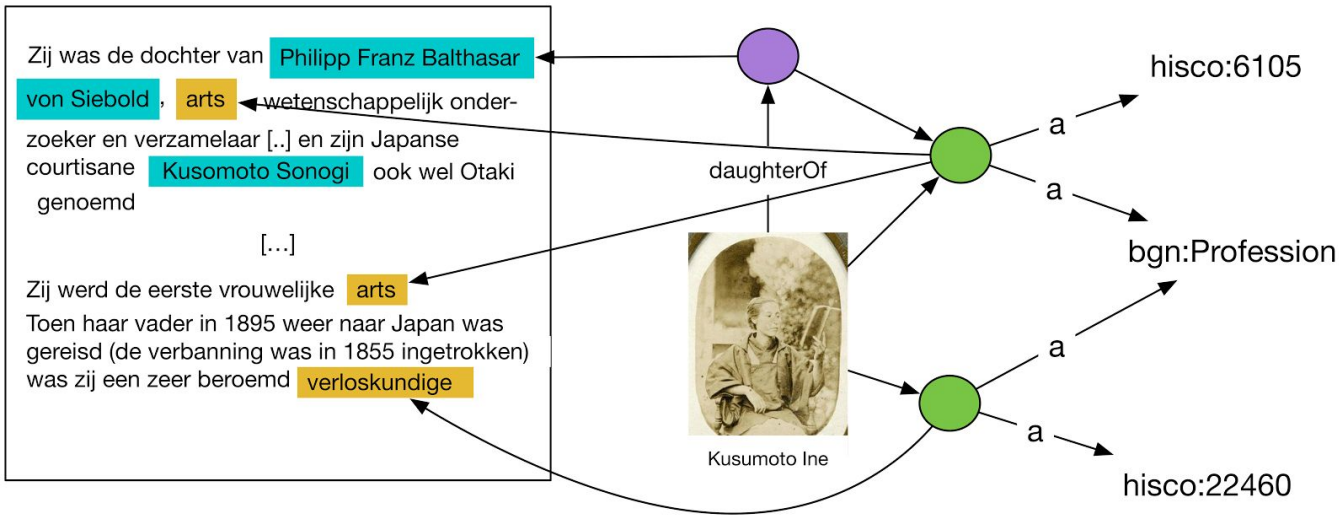


Figure 1: Analysis of Kusumoto Ine's biography. Kusumoto and her father are linked to hisco:06105 (medical doctor) and Kusumoto also to hisco:22460 (midwife).