

Abstracts DHBenelux 2017 conference

Wednesday 5 July 2017

Session A

1. TRACING TEXT TYPES IN BIBLICAL HEBREW

Wido van Peursen

Eep Talstra Centre for Bible and Computer (ETCBC), Vrije Universiteit Amsterdam

In the NWO-funded project “Tracing Syntactic Diversity in Biblical Hebrew Texts”, the Eep Talstra Centre for Bible and Computer investigates the various variables that account for the linguistic variation that can be observed in the Hebrew Bible, a collection of writings that were composed over a period of about a millennium. One of the parameters taken into account is text type.

Various types of communication show different usages of the language. The language of narratives is different from that of legal or sapiential texts. However, for a linguistic analysis a classification based on genre (“story”, “laws”) is insufficient. Genre may suggest a certain text type (e.g. a fairy tale is “narrative”), but within one text various text types may occur: in a story, the characters may use discursive text in quoted speech; in psalms or direct speech sections, a short story may develop. Sometimes the narrator addresses the listeners/readers directly and switches from narrative to discursive speech. This happens, e.g., when a fairy tale ends with *Und wenn sie nicht gestorben sind, so leben sie heute noch*.¹

In Biblical Hebrew, there is a “narrative tense” (*wayyiqtol*) similar to, e.g., the French *passé simple*, which Harald Weinrich used for distinguishing two Tempusregister: *besprechen* and *erzählen* (corresponding more or less to Émile Benveniste’s *histoire* and *discours*). Weinrich’s focus was on Roman languages. Through the work of the Semitic and Egyptian scholar Hans-Jakob Polotsky his work found an entrance in Egyptian and Semitic studies and through Ariel Shisha-Halevy, one of Polotsky’s students, also in Celtic studies. Wolfgang Schneider² introduced Weinrich’s insights into Biblical studies, where they were further developed by Eep Talstra,³ Alviero Niccacci,⁴ Gino Kalkman⁵ and others.

Building on the work of these scholars, we use the category “text type” as a feature in our linguistic database of the Hebrew Bible. We distinguish between Narrative (N), Quotation (Q) and Discursive (D). We consider text type a feature of a clause, rather than of a larger literary unit, so that we can easily handle text type shifts within one and the same literary unit. We assign the labels on the basis of syntax, rather than literary considerations. E.g.: a clause containing *wayyiqtol* is assigned the text type label N; a clause containing a vocative or a 1st or 2nd person reference the label Q; the so-called Hebrew Imperfect interrupting a narrative yields the label D, indicating those cases where the narrator directly addresses the readers.

¹ Harald Weinrich, *Tempus: Besprochene und erzählte Welt* (München: H.C. Beck, 1964).

² Wolfgang Schneider, *Grammatik des biblischen Hebräisch* (München: Claudius, 1974).

³ E.g., Eep Talstra, “Text Grammar and Hebrew Bible I”, *Bibliotheca Orientalis* 35 (1978), 168–175.

⁴ E.g., Alviero Niccacci, *The Syntax of the Verb in Classical Hebrew Prose* (Sheffield: Sheffield Academic Press, 1990).

⁵ G.J. Kalkman, *Verbal Forms in Biblical Hebrew Poetry: Poetic Freedom or Linguistic System?* (PhD dissertation, Vrije Universiteit Amsterdam, 2015).

The use of one text type within another, results in accumulative labels. Thus, a direct speech within a narrative domain receives the label NQ. Sometimes a direct speech may be quoted in another direct speech section, resulting in the label NQQ. Within a direct speech a narrative text type may appear, which introduces a short story in the mouth of one of the characters, resulting in the label NQN. Such a *Sprosserzählung* (Schneider) may again contain a direct speech, resulting in a NQNQ text type. Cf. 2 Kings 1:6.

N	They [the messengers] said (<i>wayyiqtol</i>) to him [the king]
NQ	“There came a man to meet us,
NQN	and said (<i>wayyiqtol</i>) to us
NQNQ	“Go back to the king who sent you, and say to him,
NQNQQ	“Thus says the Lord
NQNQQQ	“Is it because there is no God in Israel that you are sending to inquire of Ba’al-zebub, the god of Ekron?

In an earlier phase of the ETCBC, the text types were assigned by human researchers in interactive procedures. Recently we changed this work flow and developed algorithms for automatically assigning text types based on the above-mentioned syntactic observations. It is instructive to see the differences between the former assignment of text types in the human-computer interaction and the current automatic assignment. One such case where the automatic text type assignment leads to a different analysis occurs in Isaiah 3:14–15, where we find Q without an explicit direct speech introduction: “The Lord will bring his charge against the elders and officers of His people: “It is you...”.” Here the program assigned a Q on the basis of the 2nd person form. This had escaped the human researcher.

The automatic assignment of text types has several advantages:

- The well-defined formal criteria render the text type assignments traceable and repeatable.
- The strict application of formal criteria reveals phenomena that escape human intuition (cf. example from Isaiah 3:14–15).
- In the above-mentioned NWO project it proved to be a good starting point for investigating to what extent text type accounts for linguistic variation in the Bible.
- It makes complex textual layers and embedding visible (cf. 2 Kings 1:6)

However, there are also some challenges:

- There is the risk of circular argument based on interrelated labels such as “narrative tense” and “narrative text type”.
- The text type assignments may lead to counter-intuitive complex labels such as QNDNDN (Psalm 78:45) or NDNDN (Isaiah 9:19).
- This approach, which we inherited from the start of the ETCBC 40 years ago (when the creation of our Hebrew database started) runs the risk of being somewhat idiosyncratic, relying too much on one single study from the 1960s.

We try to parry these challenges by

- Analysing text types in interaction with other parameters, such as genre. In our statistical analysis of the distribution of linguistic phenomena in R, we take into account all other kinds of variables, as well as their possible interdependence.
- Investigating if and how we can harmonize Weinrich’s useful, but also somewhat provocative and outdated views (e.g. his denial of the expression of tense and aspect as functions of the verb)

with current insights about tense and aspect in Biblical Hebrew,⁶ and more recent studies on “Discourse Modes” and their linguistic correlates.⁷

This research provides thus an interesting case study of the interaction between linguistic theory and textual analysis, of the confrontation between research traditions within a certain discipline and a data-driven approach, and of the potential and limitations of automated analysis of ancient texts.

2. Automating genre classification of historical newspaper articles. Mapping the development of journalism’s modes of expression

Frank Harbers – University of Groningen Juliette Lonij – Dutch National Library (KB)

This paper discusses a machine learning approach to automate the genre classification of Dutch historical newspaper articles and reflects on the challenges and its value. First, we discuss how we used an existing set of metadata to create a training set for the genre classifier and the challenges we faced in connecting the metadata to the original digitized historical newspaper articles. Subsequently, the paper outlines a machine learning approach to predict the genre of a newspaper articles, discussing and evaluating the different tools that were tested in the process⁸. Finally, it reflects on the way a traditional rule-based approach to determining genre relates to a machine learning approach.

Examining genre

Defined as “language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms” (Handford 2010), genre can elucidate the underlying goals, norms and practices of journalism as a discourse. Examining journalistic genres from a historical perspective therefore elucidates how newspapers’ conception of journalism developed. Yet, this type of longitudinal textual research is highly time consuming and still scarce. Moreover, the few attempts to systematically examine newspaper material, using social scientific methods such as quantitative content analysis, still only cover a fraction of the available material (Broersma, 2011; Harbers, 2014).

Automating such content analyses would be highly beneficial for research into the discursive development of newspaper journalism. This paper therefore critically discusses an approach to automate genre classification. This is a daunting task as genres are dynamic and can change or fade away over time while new ones can emerge. Moreover, genres are ideal types, which means the textual manifestations do not always match all the characteristics perfectly, nor can they always be clearly delineated from other genres.

⁶ E.g., Jan Joosten, *The Verbal System of Biblical Hebrew. A New Synthesis Elaborated on the Basis of Classical Prose* (Jerusalem Biblical Studies 10; Jerusalem: Simor, 2012).

⁷ E.g., Carlota S. Smith, *Modes of Discourse. The Local Structure of Texts* (Cambridge Studies in Linguistics 103; Cambridge: Cambridge University Press, 2003).

⁸ The source code for training the classifier and applying it to new examples is available on GitHub (<https://github.com/jlonij/genre-classifier>) and everybody can experiment with the classifier through a graphical web interface created at <http://www.kbresearch.nl/genre>. This dataset was the result of a large-scale research project into the historical development of European newspapers with the title ‘Reporting at the boundaries of the public sphere. Form, Style and Strategy of European Journalism, 1880-2005’.

A machine learning approach to automatically classify genre

Building on an existing set of manually coded metadata, describing several textual characteristics,

such as genre, of a large sample (33.000) of historical newspaper articles², this paper outlines a machine learning approach to automate the genre classification of historical newspaper articles. This dataset thus provided us with metadata about a number of historical articles that was used to train and formally evaluate a classifier that is able to automatically predict the genre of additional samples of historical newspaper articles. Yet, the existing metadata needed to be linked to the corresponding digitized articles in the digital newspaper archives of the KB.

The paper will first discuss this linking process. We first selected the most promising candidate links for each item in the original data set, based on the position of the article on the page, its size, and the presence of images and quotes. A simple classifier was then trained to select the best link from the candidate set, if any, based on more precise features such as the size difference between the article and the candidate, as well as author mentions and subject matter. By only accepting links predicted with a relatively high confidence value approximately 50% of all articles could be automatically linked, with an error rate of 0.5%.

Subsequently, we will outline and discuss how the resulting data set was used to train the actual genre classifier. After the articles were pre-processed with the Natural Language Processing suite 'Frog', the annotated texts were examined for their textual features, including the length of the article, the number of direct quotes, the number of adjectives, various types of pronouns, and the number and position of named entities in the text. The selection of these features is based on the genre definitions of the codebook of the manual content analysis.

These features were used to train a classifier to choose one of eight possible genres for each article, ranging from news report to opinion article. We evaluated the performance through 10-fold cross-validation, using stratified sampling to create relevant subsets. A linear SVM classifier was chosen after comparison of various evaluation metrics with a number of other options (Naïve Bayes, non-linear SVMs and some simple neural networks), yielding the best results with an accuracy of 65%. It is important to note here that human coders do not always agree on what the right genre is. The intercoder agreement for genre in the manual content analysis was around 80% (Krippendorff's alpha, taking into account chance, was between 0.7 and 0.8 in different groups of coders). As such, 65% is considered a very promising result.

Finally, we reflect on the relation between a rule-based and machine learning approach to the classification of genre. We will discuss the significance of individual features in the machine learning process and show how the 'confusion matrix' provides valuable information about the common mistakes of the classifier and which genres are most difficult to predict. Moreover, as the probability for the predicted genre as well as for the other genres is known, we will discuss how these numbers offer insights in the dynamic nature of journalistic genres.

Bibliography

- Broersma, M. (2011). 'Nooit meer bladeren. Digitale krantenarchieven als bron'. In: *Tijdschrift voor Mediageschiedenis* 14(2): 29-55
Handford, M. (2010). 'What can a corpus tell us about specialist genres'. In: 'o Keeffe, A. & McCarthy, M. (eds.), *The Routledge Handbook for Corpus Linguistics*. New York: Routledge.
- Harbers, F. (2014). *Between Personal Experience and Detached Information. The Development of Reporting and the Reportage in Great Britain, the Netherlands and France, 1880-2005*. PhD University of Groningen

3. Generating Interactive Narratives from Wikipedia Articles

Keywords: Computational Creativity, Interactive Narratives, Wikipedia, Chatbots

Ben Burtenshaw
Computational Linguistics & Psycholinguistics Research Center
Universiteit Antwerpen
benjamin.burtenshaw@uantwerpen.be

Tom De Smedt
Experimental Media Research Group
Sint Lucas Antwerpen School of Arts
info@emrg.be

Mike Kestemont
Computational Linguistics & Psycholinguistics Research Center
Universiteit Antwerpen
mike.kestemont@uantwerpen.be

Stories play a vital role in the lives of children. The alternative worlds they produce encourage imagination and creativity, but also transform knowledge into structures that children can understand and relate to. We present an interactive story system that creates narratives from Wikipedia articles, and reveals them through dialogue with a user. Using state-of-the-art narrative generation tools and a chatbot dialogue system, information from Wikipedia is revealed to the child based on their input. Generating narratives from any text has long been a goal of Artificial Intelligence researchers because narrative structures are useful for learners of all ages. However, many of the existing story generation systems have relied on hand-written techniques that cannot meet the scale of data online. In recent years search-based systems have been able to incorporate broader topics, but they have sacrificed continuity, producing fragmented narrative. Here we propose a search-based system that scours Wikipedia for articles relating to user input, and then restricts its generation material to that article; in doing so, the system utilises the defined topic and chronology of the article to retain context.

Narrative generation has been a central topic within the field of computational creativity for decades. One of the first examples is *Tale-Spin*, a system that generates Aesop's Fables guided by the user's input (Meehan 1977). *Tale-spin* produces an innovative form of interface; however, it struggles to deal with undirected user input. *Universe* by Michael Lebowitz draws on a database of character definitions, plot outlines, and dialogues, to weave together new stories; however, the system has a tendency to become repetitive due to its limited content. Callaway and Lester produced *Storybook*, a narrative prose generator that identifies the temporal markers in a text, and uses them to generate a new narrative text (Callaway and Lester 2002). Systems of this kind tend to reproduce narrative tropes, and ultimately become boring (Swanson and Gordon 2008). McIntyre and Lapata developed one of the first search-based narrative systems, that uses user-input to search a database for relating phrases (McIntyre and Lapata 2009). This architecture produces interesting relationships between phrases, but is too sporadic to create a meaningful narrative. To counter these semantic fluctuations, Riedl and Bulitko use dialogue to guide the generated text (Riedl and Bulitko 2012). However, by their own admission, this becomes monotonous to the user, with little room for surprise.

We propose a dialogue system that takes input from the user and generates a story in relation to a relevant Wikipedia article. This uses a search-based retrieval approach similar to previous systems (Riedl and Bulitko 2012; McIntyre and Lapata 2009); however, we take advantage of Wikipedia for topic and contextual grounding. For example, by taking the proper nouns in a template sentence, and replacing them with those from a Wikipedia article, the system produces a story featuring places and

characters from history: The user might ask “Why did the Egyptians have pyramids?”, the system would use key points in the Wikipedia article ‘Pyramids of Ancient Egypt’, to assemble a set of narrative concepts, and query the user to see if they are interested in those topics.

We have trained the system on children’s stories, which means it uses a machine learning approach to select responses that are most similar to those stories. A dialogue is started by inputting a topic which is used to retrieve a complete Wikipedia page. The page structure guides the dialogue sequence, and the language is used to create responses. First the system defines a set of narrative concepts based on the sections of the wikipedia page, then it uses grammar based techniques to create phrases for each concept. The narrative concepts are vector based representation of each page section, which are used to compare the most important strings from a section. Dialogue begins when the system has built a database of possible responses that are stored in a database. As the dialogue progresses the core phrases are added to and removed; in effect, acting as a narrative context. Dialogue is facilitated by a Markov Decision Process that matches the user’s input to possible responses based on a reward function from training. Dialogue history is also added to the database, which means the system learns from the dialogue itself, and also avoids repetition.

Using named entity recognition to add names and places grounds the narrative of the story. This practice draws on pedagogical theories of inquiry, that propose children learn efficiently through self initiated requests for information (Conle 2000; Mcquiggan et al. 2008). Children are able to construct this information into their own story, which in turn fortifies that knowledge. In practice, a narrative system like this would need to be contextualised for the child; this could be a fictional characterisation of the system as naive and in need of guidance. For example, the forgetful robot that needs help to explain their story.

Narratives are a proven part of education, and generating them autonomously is a long-term aim of artificial intelligence. Here we speculate upon a contained usage of narrative generating technology, which uses the structure of Wikipedia to rearticulate text in a form suitable for children. Over the next four years our aim is to develop a general story generation system for children. At DHBenelux we will present a working prototype of this system.

References

- Callaway, C., and J. Lester. 2002. ‘Narrative Prose Generation’. *Artificial Intelligence* 139 (2): 213–252.
- Conle, Carola. 2000. ‘Narrative Inquiry: Research Tool and Medium for Professional Development’. *European Journal of Teacher Education* 23 (1): 49–63. doi:10.1080/713667262.
- McIntyre, Neil, and Mirella Lapata. 2009. ‘Learning to Tell Tales: A Data-Driven Approach to Story Generation’. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, 217–25. Association for Computational Linguistics.
- Mcquiggan, Scott W, Jonathan P Rowe, Sunyoung Lee, and James C Lester. 2008. ‘Story-Based Learning: The Impact of Narrative on Learning Experiences and Outcomes’. In *International Conference on Intelligent Tutoring Systems*, 530–539. Springer Berlin Heidelberg.
- Meehan. 1977. ‘TALE-SPIN, An Interactive Program That Writes Stories’. In *5th International Joint Conference on Artificial Intelligence*, 91–98.
- Riedl, Mark O., and Vadim Bulitko. 2012. ‘Interactive Narrative: an Intelligent Systems Approach’. *Ai Magazine* 34 (1): 67.
- Swanson, R. and Gordon, A. 2008. ‘Say Anything: A Massively Collaborative Open Domain Story Writing Companion’. In *Proceedings of the 1st International Conference on Interactive Digital Storytelling. Lecture Notes in Computer Science. Vol. 5334*. Berlin: Springer.

1. Linking multi-disciplinary data sources for a historical research platform

Kalliopi Zervanou¹, Wouter Klein², Peter van den Hooff², Frans Wiering¹ and Toine Pieters²

¹Information & Computing Sciences Department, Utrecht University

²Freudenthal Institute, History & Philosophy of Science, Utrecht University

The problem of *information access* is a challenge in making digitised data sources available. Historians need to identify information in digital material pools, scattered across collections and often lacking semantic links to a topic of interest. This problem is addressed by the development of various collection-specific metadata schemas, such as MARC 21 (Library of Congress, 2010), and generic metadata schemas, such as the Dublin Core Metadata Initiative (DCMI, 2011). Moreover, diverse metadata schemas are mapped to each other (Bountouri and Gergatsoulis, 2009), or to custom (Liao et al., 2010) or standard ontologies (Lourdi et al., 2009), such as the CIDOC Conceptual Reference Model (CIDOC, 2006). A dominant trend in recent approaches is the *linked-data* approach (Berners-Lee, 2006; Bizer et al. 2009). Besides information access, the amount and the complexity of information accessible gives rise to an *information presentation* challenge, whereby data overviews should highlight interesting data aspects requiring detailed inspection. Additionally, for digital methods to support collaborative research, the problem of *information validation and sharing* must be addressed. This issue calls for transparency of data and methods and reproducibility of results, or verification of the arguments made. It also entails validation of computational results and algorithmic processes determining information access, in such a way that the eventual data or system limitations and biases are known and the processes are trustworthy and verifiable.

In our work, we address these challenges of information access, presentation, validation and sharing from a twofold research perspective:

- I. Integration and semantic linking of existing, multidisciplinary data sources;
- II. Development of a research platform that supports access, presentation, validation and sharing of complex, interlinked data.

Our particular domain of application relates to the history of botanical drug components from the New World in the early modern period (17-18th century). More specifically, it concerns highlighting phenomena denoting developmental processes of remedies or *drug trajectories*, such as the evolution of economic importance, ethical attitudes, scientific interests, trade and knowledge circulation (Gijswijt-Hofstra et al., 2002; Pieters, 2004; Friedrich and Müller-Jahncke, 2009; Klein & Pieters, 2016).

For this purpose, we integrate sources comprising of pharmaceutical data, such as the *Pharmaceutical Historical Thesaurus* (Klein & van den Hooff, 2013), archaeobotanical data, such as *RADAR* (van Haaster & Brinkkemper, 1995; RCE, 2013), botanical data, such as the *National Herbarium of the Netherlands* (Creuwels, 2014), the *Economic Botany database* (Hoffman, 2011) and the *Snijpendaal Catalogue database* (van Reenen, 2007), colonial trade data, such as the database of the accounting books (*Boekhouder-Generaal*) of the Dutch East India Company (Schooneveld-Oosterling et al., 2013) and linguistic dictionaries, such as the Chronological Dictionary of Dutch (van der Sijs, 2001).

A notable recent approach to the issue of digital formats integration is the one adopted in the Timbuctoo infrastucture (Andersen, 2013). Most approaches opt for conversion to a recommended metadata schema, such as SKOS (Miles & Bechhofer, 2009), or a common data model such as the Europeana Data Model (EDM, 2016). However, apart from the diversity of digital formats an

important aspect in integration lies in the reuse and re-purposing of resources originally built for a different audience and purpose.

In our approach, integration entails concept mapping, not only across disciplines, but also in time. Thus, data source integration calls for support for the evolution of science from the 16th century onwards to re-classify and re-define concepts. Additionally, it entails dealing with phenomena of historical term variation and ambiguity which gradually give way to spelling standardisation and current nomenclature conventions in e.g. botany and biology. Furthermore, we account for under-specificity and ambiguity of information found in historical sources while maintaining associations with potentially related concepts and context. Most importantly, we provide references for information provenance tracing and validation. For these purposes, we resort to designing our own ontology, where e.g. ambiguous terms are connected to multiple concepts, temporal periods and reference sources, and where mappings are provided across essential historical and current taxonomies. Our data sources are semi-automatically enriched with additional information, such as geographical coordinates and named entities. Moreover, inconsistencies within and across data sets are semi-automatically identified and normalised. Finally, data sources are integrated following a linked data approach allowing for extensions to other linked open data and eventually capitalising on techniques such as reasoning, which may extend explicit information in our data sets with implicitly inferred information.

Our *Time Capsule* research platform⁹ implements our solutions to information access, presentation and validation challenges. It is a scalable working platform currently querying more than 55 million RDF triples. It is often difficult for a non-expert user to perform queries, either because they are unfamiliar with the required terminology, or because they are unfamiliar with the underlying data model. Our solution to this issue lies in providing two querying strategies, one that supports a faceted, exploratory, guided search and browsing of information by means of links, photos, and keyword auto-completion suggestions and one that supports the creation of ad hoc queries. Our exploratory search mode is intended to engage a wider audience and reveal to both experts and non-expert users the underlying data content and structure. Ad hoc queries are in essence ad hoc RDF SPARQL queries (Prud'hommeaux & Seaborne, 2008) to our data. However, given that most users are neither familiar with SPARQL, nor with the content and structure of our datastore, a query wizard is provided that assists users in forming natural language queries, such as *"Which drug component(s) are made out of the plant Acorus calamus L. and which parts of the plant were used?"*

Search results are presented as an overview of all available information on the query topic and users may "zoom-in" on specific information by following links that provide more detailed geographical, temporal and concept relation visualisations. Such visualisations are mainly intended to provide overviews in the evolution of phenomena related to drug trajectories, such as for instance change in a plant part used as medical ingredient, trade routes of botanical products, or geographical distributions in time of known concepts in Latin/scientific terms vs. lay terms, the latter indicating public knowledge and familiarity with a given plant or drug.

References

- Andersen, J. A., Filarski, G. J., Haentjens Dekker, R., Maas, M. & Ravenek, W. (2013).** Timbuctoo data repository infrastructure (version 1.0), Huygens ING – ICT, Amsterdam, The Netherlands.
- Berners-Lee, T. (2006).** Linked Data. Document version: June 2009. In: Design Issues, W3C. Available online at: <https://www.w3.org/DesignIssues/LinkedData.html>

⁹ Time Capsule system: <http://timecapsule.science.uu.nl/timecapsule/#/login> Logging in as a *Guest* allows full access to the system functionality except saving your search results.

- Bizer, C., Heath T. and Berners-Lee T. (2009).** Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, vol. 5(3), pp. 1-22. DOI: 10.4018/jswis.2009081901
- Bountouri L. and Gergatsoulis M. (2009).** Interoperability between archival and bibliographic metadata: An EAD to MODS crosswalk. *Journal of Library Metadata*, 9(1-2):98–133.
- CIDOC (2006).** The CIDOC Conceptual Reference Model. CIDOC Documentation Standards Working Group, International Documentation Committee, International Council of Museums. Available online at: <http://www.cidoc-crm.org/>.
- Creuwels, J. (2014).** The National Herbarium of the Netherlands. Naturalis Biodiversity Center, Leiden. Available online at: <http://herbarium.naturalis.nl/>
- DCMI (2011).** The Dublin Core Metadata Initiative. Available online at: <http://dublincore.org/>.
- EDM (2016).** Europeana Data Model – Mapping Guidelines v2.3, 18 November 2016, Europeana Network Association. Available online at: <http://pro.europeana.eu/page/edm-documentation>
- Friedrich, C. and Müller-Jahncke, W.-D. (eds.) (2009).** Arzneimittelkarrieren: zur wechsellvollen Geschichte ausgewählter Medikamente: die Vorträge der Pharmaziehistorischen Biennale in Husum vom 25-28. April 2008, Stuttgart: Wissenschaftliche Verlagsgesellschaft.
- Gijswijt-Hofstra, M., Van Heteren, G. M. and Tansey, E. M. (eds.) (2002).** Biographies of remedies: drugs, medicines and contraceptives in Dutch and Anglo-American healing cultures. *Clio medica* 66, Amsterdam: Rodopi.
- van Haaster H. and Brinkkemper O. (1995).** RADAR, a Relational Archaeobotanical Database for Advanced Research. *Vegetation History and Archaeobotany*, vol. 4(2), pp. 117-125, Springer.
- Hoffman, B. (2011).** The Naturalis Economic Botany database. Naturalis Biodiversity Center, Leiden.
- Klein, W. and Pieters, T. (2016).** The Hidden History of a Famous Drug: Tracing the Medical and Public Acculturation of Peruvian Bark in Early Modern Western Europe (c. 1650–1720). *Journal of the History of Medicine and Allied Sciences*, Vol. 71(4), pp. 400–421. DOI: 10.1093/jhmas/jrw004
- Klein, W. and van den Hooff, P. C. (2013).** Farmaceutische Historische Thesaurus. National Museum for the History of Pharmacy, Utrecht.
- Liao, S.-H., Huang, H.-C., and Chen, Y.-N. (2010).** A semantic web approach to heterogeneous metadata integration. In: *Proceedings of ICCCI '10, LNCS vol. 6421*, pp. 205–214, Kaohsiung, Taiwan. Springer.
- Library of Congress (2010).** MARC standards. Network Development and MARC Standards Office, Library of Congress, USA. Available online at: <http://www.loc.gov/marc/index.html>.
- Lourdi, I., Papatheodorou C., and Doerr M. (2009).** Semantic integration of collection description: Combining CIDOC/CRM and Dublin Core collections application profile. *D-Lib Magazine*, 15(7/8).
- Miles, A. and Bechhofer S. (eds) (2009).** SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 18 August 2009. Available online at: <http://www.w3.org/TR/skos-reference>
- Pieters, T. (2004).** Historische trajecten in de farmacie: medicijnen tussen confectie en maatwerk. Inaugural lecture – Hilversum.
- Prud'hommeaux, E. and Seaborne A. (eds.) (2008).** SPARQL Query Language for RDF. W3C Recommendation, 15 January 2008. Available online at: <https://www.w3.org/TR/rdf-sparql-query/>

RCE (2013). RADAR, a Relational Archaeobotanical Database for Advanced Research. Rijksdienst voor het Cultureel Erfgoed, Ministerie van Onderwijs, Cultuur en Wetenschap. Available online at: <https://archeologiein nederland.nl/bronnen-en-kaarten/radar>

van Reenen, G. (2007). Snippendaalcatalogus database. Hortus Botanicus Amsterdam. Available online at: <http://dehortus.nl/en/Snippendaal-Catalogue>

Schooneveld-Oosterling, J., Knaap, G., Karskens, N., Smit-Maarschalkerweerd, D., Tetteroo, S., van den Tol, J., Nijhuis, H., van Wijk, K., Kunst, A., Buijs, J., Jongma, M., Boer, R. (2013). Boekhouder-Generaal Batavia. Huygens ING. Available online at: <http://resources.huygens.knaw.nl/boekhoudergeneraalbatavia>

van der Sijs, N. (2001). Chronologisch Woordenboek. Available online at: http://dbnl.org/tekst/sijs002chro01_01/

2. A Linked Data Approach to Disclose Handwritten Biodiversity Heritage Collections

Lise Stork, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
l.stork@liacs.leidenuniv.nl

Andreas Weber, Department of Science, Technology and Policy Studies (STePS), University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
a.weber@utwente.nl

Over the last decade, natural history museums in and beyond the Netherlands have heavily invested in digitizing and extracting biodiversity information from manuscript and specimen collections (Heerlien et al. 2015; Pethers and Huertas, 2015; Svensson, 2015). In particular handwritten fieldnotes describing occurrences of species in nature (see illustration) form an important but often neglected starting point for researchers interested in long-term habitat developments of a specific area and the history of scientific ordering, writing and collecting practices (Blair 2010; Bourget 2010; Eddy 2016). In order to disclose handwritten descriptions of flora and fauna and related specimen and drawings collections, natural history museums usually resort to manual enrichment methods such as full text transcription or keyword tagging (Ridge 2014; Franzoni et al. 2014). Often these methods rely on crowdsourcing, where online volunteers annotate pages with unstructured textual labels (Field Book Project 2016). More recently, curators of archives, data scientists and historians have started to experiment with semi-automatic annotation systems for historical manuscript collections such as the MONK system (Schomaker et al. 2016). Since MONK is a supervised learning system, a large amount of properly recognized textual labels is necessary to safeguard the system's recognition abilities.



Thus, although such practices have the potential to yield high quality data, merely annotating pages with unstructured textual labels raises two problems: First, without suggestions driven by semantic

knowledge, it will be hard for volunteers or a machine to start annotating handwritten pages. Not only in the context of our case study, which deals with fieldnotes written in early nineteenth century insular Southeast Asia, but also in the context of other manuscript collections, one needs a thorough knowledge of paleography, and historical and taxonomic background information (Causer and Terras 2014). Semantics can aid the annotation process when dealing with ambiguity or provide suggestions in cases where words are hard to read and too little example instances are available. For instance, when a fieldnote describes an expedition in East-Java, a species of frogs of West-Celebes can be ruled out. Second, unstructured textual annotation will eventually result in an inefficient search process on the side of the user. Traditional keyword-based search leads to many irrelevant results or requires specific prior knowledge regarding the content. To answer more general and expressive queries, semantic relations between annotations need to be considered as well (Elbassuoni, et al. 2010).

In order to help solve such problems this paper argues for the development and application of a semantic model for semi-automatic semantic annotation. The model aggregates existing metadata standards and ontologies, following the Linked Data principles, and prepares them for semantically annotating and interpreting the *Named Entities (NEs)* in the fieldnotes of digitized natural historical collections.¹⁰

The case study of this paper is a collection of 8000 fieldnotes gathered by the Committee for Natural History of the Netherlands Indies (*Natuurkundige Commissie voor Nederlandsch-Indië*, further referred to by the acronym NC). In the first half of the nineteenth century, naturalists of the NC charted the natural and economic state of the Indonesian Archipelago and returned a wealth of scientific observations which are now stored in the archives and depot of Naturalis Biodiversity Center in Leiden (Mees 1994; Klaver 2007). An in-depth historical analysis reveals that Heinrich Kuhl (1797-1821), Johan Coenraad van Hasselt (1797-1823) and other travelers of the NC use the following NEs to structure their fieldnotes (see illustration displaying a bundle of NC fieldnotes) while traveling in insular Southeast Asia: collecting localities, dates, collectors' names, taxonomic names, and references to other printed or handwritten sources. Kuhl and Van Hasselt, for instance, regularly use the illustrations of printed works such as the *Voyage de découvertes aux terres australes (1807-1816)* by M.F. Péron as visual point of reference for their fieldnote descriptions. While links to published resources can be easily established by linking them to domain specific repositories of digitized books such as the Biodiversity Heritage Library (BHL), collection localities, taxonomic names and collectors' names are more difficult to process.



In order to be able to identify, annotate and interlink such NEs in a semi-automatic way, this paper proposes the implementation of a Knowledge Base (KB). The KB has two goals: first, the underlying data structure of the KB enables cross-matching of resources within and across fieldnote

¹⁰ The project Semantic Blumenbach thinks in a similar direction, but then with a focus on published material (Wettlaufer et al. 2015).

collections. In order to realize this function a lightweight application ontology written in RDF¹¹ and OWL¹² is suggested that serves as a schema to semantically structure the KB. It expresses species observations, ensures their provenance in relation to the digitized fieldnotes and builds on existing metadata and ontology standards. Entities in turn are described using uniform resource identifiers (URIs). This allows for an integration of the fieldnote annotations into the web of Linked Data (LD) and ensures interoperability with other digital collections (Hallo et al. 2016). Second, the logical characteristics of the properties in the ontology enable a reasoner system to suggest possible NEs. In order to provide possible labels regarding these NEs, the KB is prepopulated with lists extracted from thesauri, gazetteers, and taxonomies. As regards collection localities we, for instance, draw upon the *GEOnets Names Server (GNS)*, a large semantically structured database containing historical and present-day geographical locations in insular Southeast Asia. Biological species names can be drawn from the Linnaean taxonomy of species which was already well established at the time of the NC (Farber 2000; Beckman 2012). As regards person names we rely on the database *Cyclopedia of Malaysian Collectors* which M. J. van Steenis-Kruseman compiled in the 1960s and 1970s.¹³ Taken together, by prompting users to annotate with terms from the KB, a semantic network of annotations is formed that is able to improve the quality of the annotations and bootstraps the annotation process. The ontology and an implementation of the KB based on our case study, together with possibilities regarding supported querying and reasoning techniques, will be discussed in more detail during the presentation.

Bibliography

- Beckman, J. "The Swedish Taxonomy Initiative : Managing the Boundaries of 'Sweden' and 'Taxonomy'" In *Scientists and Scholars in the Field: Studies in the History of Fieldwork and Expeditions*, edited by K.H. Nielsen, H. Harbsmeier, and Ch. J. Ries, 395–414. Aarhus: Aarhus University Press, 2012.
- Bourguet, M.-N. "A Portable World: The Notebooks of European Travellers (Eighteenth to Nineteenth Centuries)." *Intellectual History Review* 20, no. 3 (2010): 377–400.
- Causser, T. and M. Terras. ""Many Hands Make Light Work. Many Hands Together Make Merry Work": Transcribe Bentham and Crowdsourcing Manuscript Collections."" In *Crowdsourcing Our Cultural Heritage*, 57–88. Surrey: Ashgate, 2014.
- Eddy, M. D. "The Interactive Notebook: How Students Learned to Keep Notes during the Scottish Enlightenment." *Book History* 19, no. 1 (2016): 86–131.
- Elbassuoni, S., Ramanath, M., Schenkel, R., and Weikum, G. "Searching RDF Graphs with SPARQL and Keywords". *IEEE Data Eng. Bull.*, 33(1), (2010), 16-24.
- Farber, P.L. *Finding Order in Nature: The Naturalist Tradition from Linnaeus to E.O. Wilson*. Baltimore, Md.: Johns Hopkins University Press, 2000.
- Field Book Project, Smithsonian National Museum of Natural History: <http://naturalhistory.si.edu/fieldbooks/> [accessed 15 February 2017].
- Franzoni, Ch. and H. Sauermann, "Crowd science: The organization of scientific research in open collaborative projects," *Research policy* 43, no. 1 (2014), 1-20.

¹¹ <https://www.w3org/RDF/> [accessed February 15, 2017].

¹² <https://www.w3org/OWL/> [accessed February 15, 2017].

¹³ The database is available online: <http://www.nationaalherbarium.nl/FMCollectors/> [accessed February 15, 2017]

GEONets Name Server, <http://geonames.nga.mil/gns/html/> [accessed February 15, 2017]

Hallo, M., et al. "Current state of Linked Data in digital libraries." *Journal of Information Science* 42.2 (2016): 117-127.

Heerlien, M., J. Van Leusen, S. Schnörr, S. De Jong-Kole, N. Raes, and Kirsten Van Hulsen. "The Natural History Production Line: An Industrial Approach to the Digitization of Scientific Collections." *J. Comput. Cult. Herit.* 8, no. 1 (February 2015): 3:1–3:11.

Klaver, Ch.J.J. *Inseparable Friends in Life and Death: The Life and Work of Heinrich Kuhl (1797-1821) and Johan Conrad van Hasselt (1797-1823), Students of Prof. Theodorus van Swinderen*. Groningen: Barkhuis, 2007.

Mees, G.F. and C. van Achterberg. "Vogelkundig onderzoek op Nieuw Guinea in 1828: terugblik op de ornithologische resultaten van de reis van Zr. Ms. Korvet Triton naar de zuidwest kust van Nieuw-Guinea." *Zoologische Bijdragen* 40 (1994): 3–64.

Péron, F., N. Baudin, L.C. Desaulses de Freycinet, Ch. Alexandre Lesueur, and N.-M. Petit. *Voyage de Découvertes Aux Terres Australes* (Paris : De l'Imprimerie impériale, 1807).

Pethers, H. and B. Huertas. "The Dollmann Collection: A Case Study of Linking Library and Historical Specimen Collections at the Natural History Museum, London." *The Linnean* 31, no. 2 (2015): 18–22.

Ridge, M. (ed.), *Crowdsourcing our cultural heritage* (Ashgate: Farnham, 2014).

Schomaker, L., A. Weber, M. Thijssen, M. Heerlien, A. Plaat, S. Nijssen, et al. "Making Sense of Illustrated Handwritten Archives." In *Book of Abstracts, Digital Humanities Conference 2016 Krakow*, 764–66, 2016.

Svensson, A. "Global Plants and Digital Letters: Epistemological Implications of Digitising the Directors' Correspondence at the Royal Botanic Gardens, Kew." *Environmental Humanities* 6 (2015): 73–102.

Wettlaufer, J., Ch. Johnson, M. Scholz, M. Fichtner, and S. Ganesh Thotempudi. "Semantic Blumenbach: Exploration of Text–Object Relationships with Semantic Web Technology in the History of Science." *Digital Scholarship in the Humanities* 30, Suppl. 1 (December 1, 2015): 187–98.

3. Linked cultural events: Digitizing past events and its implications for analyzing and theorizing the 'creative city'

Harm Nijboer (Huygens ING)

Claartje Rasterhoff (University of Amsterdam)

Introduction

This paper introduces 'linked cultural events' as a novel methodological framework that allows for the systematic analysis of cultural expressions in their urban context. The events-based approach is inspired by datasets developed in the research program *CREATE: Creative Amsterdam: An E-Humanities Perspective* (University of Amsterdam, 2014-present).¹⁴ In this program, the cultural sectors of performing arts take up a particularly prominent position, as data on for instance music, theatre and cinema programming is available in various formats. In terms of methodology, the data

¹⁴ www.create.humanities.uva.nl.

on performing arts allows us to move beyond biographical data on producers, and develop a methodological framework in which different data types can be studied in conjunction. The framework of LCE has two main characteristics: 1) it posits cultural events as analytical units with structural properties and linkages to actors, institutions and urban properties (linked cultural events); and 2) it is connected to a data structure which allows for querying the connections between these units of analysis (linked data). In this paper, we discuss how the concept of 'linked events' can be used to map and analyse urban cultural life.

Events and the city

Studies in the social sciences and humanities research offer valuable insights in conditions and mechanisms favorable to creativity and innovation, emphasizing for instance the role of agglomeration, and labour mobility and diversity.¹⁵ Little to no attention is being paid to what actually makes cities come to life: the cultural expressions themselves, and in particular events such as exhibitions, concerts, plays, and publications. Recent historical research has, furthermore, emphasized the limits of such generalizations on sources of creativity, stressing the importance of time- and place- specific characteristics and circumstances.¹⁶

The events-based approach may help to address some of these issues. Much has been written about how events should be conceptualized and about the role of events in studying and writing history.¹⁷ Moreover, theoretical and conceptual thinking about events is not limited to historiography but expands to the fields of action theory in philosophy and social theory.¹⁸ Events also feature as devices in structuring heritage data and as building blocks for online reconstructions of historical narratives.¹⁹ Datasets of events over time have, moreover, been used in the broader field of data analytics, for instance in event-based network analyses, to add temporality and dynamism to otherwise static information systems.²⁰ Building on insights from these different lines of research, we emphasize that networks of events should also be considered as units of analysis.

Linked cultural events

A large number of contemporary social theorists rejects the notion of a cultural act or event as an expression (or representation) of a given culture. Instead culture should be understood as a collection of performative acts or events.²¹ By performativity we mean that an event calls or recalls something (a piece of art, a cultural code or trait) into being. A play, for instance, must be performed (staged, read, remembered) to be there. Events, moreover, do not occur in isolation. Each event

¹⁵ Cf. Marjatta Hietala Peter Clark, 'Creative Cities', in: *The Oxford Handbook of Cities in World History*, edited by Peter Clark. Oxford: Oxford University Press 2013.

¹⁶ Ilja Van Damme and Bert De Munck (eds.), *Creative Cities 1500-2000. The Historical Fabrication of Cities as Agents of Economic Innovation and Creativity*, London: Routledge forthcoming.

¹⁷ Cf. Ryan Shaw, 'A Semantic Tool for Historical Events', *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Atlanta, Georgia, 14 June 2013: 38–46; W.H. Sewell, 'Historical events as transformations of structures: Inventing revolution at the Bastille', *Theory and Society* 1996, 25: 841-881.

¹⁸ Roberto Casati & Achille Varzi, "Events", in: Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), <http://plato.stanford.edu/archives/win2015/entries/events>.

¹⁹ Cf. Victor de Boer, Johan Oomen, Oana Inel, Lora Aroyo, Elco van Staveren, Werner Helmich, Dennis de Beurs, 'DIVE into the event-based browsing of linked historical media', *Journal of Web Semantics*, 35/3: 152-158.. DOI: 10.1016/j.websem.2015.06.003. See also: <http://www.ehumanities.nl/events-working-group>.

²⁰ E.g. Joshua O'Madadhain, Jon Hutchins, Padhraic Smyth (2005), 'Prediction and Ranking Algorithms for Event-based Network Data', *SIGKDD Explor. Newsl.*, 7(2), pp. 23-30, [doi:10.1145/1117454.1117458](https://doi.org/10.1145/1117454.1117458).

²¹ Peter Dirksmeier & Ilse Helbrecht (2008), 'Time, Non-representational Theory and the "Performative Turn"—Towards a New Methodology in Qualitative Social Research', *Forum: Qualitative Social Research* 9 (2), pp. 1-24. <http://www.qualitative-research.net/index.php/fqs/article/view/385/839>.

involves the actions and/or presence a number of entities. These entities can be human agents (e.g. performers and spectators), non-human agents (organizations), material objects (places, artifacts, etc.) or immaterial objects (concepts, code). And these entities are in their turn likely to be involved in other events as well. Already in 1964 the ethnolinguist Dell Hymes defined 'communities' as 'systems of communicative events'.²² In order to operationalize this interpretation for historical research, we conceptualize cultural communities as webs of linked cultural events (LCEs).

LCE's thus form an infrastructure for combining, analyzing, and visualizing existing cultural datasets in a network that exposes their relations and interdependencies, and that allows for quantitative analysis. On the level of advanced data handling, the LCE approach has a strong affinity with Semantic Web technology and the associated Linked Open Data paradigm which have evolved in leading principles in the handling of historical and cultural heritage data in recent years. Notwithstanding the limitations and complexities of Semantic Web technology, the great practical advantage of this technology is that it enables us to connect single resource data to external resources. This is not only important for our understanding of cultures as webs as LCEs, but proves to be inherent to the data on for instance theatre and concert programs.

Visualising and analyzing LCE's

In the final section of the paper we will use two recently developed datasets to present analyses of linked cultural events. ONSTAGE (Online Datasystem of Theatre in Amsterdam in the Golden Age) contains information on the repertoire, performances, popularity and revenues of the cultural program in Amsterdam's public theatre during the period 1637 - 1772.²³ The FELIX: Felix Meritis Programming Database stores and links data on concerts held in the famous Amsterdam concert hall Felix Meritis between 1832 and 1888.²⁴ In these datasets linkages have been created to, for instance, genre characteristics and biographical data in external international resources such VIAF (Virtual International Authority File) and the data branches of the Wikipedia family (DBpedia and Wikidata). By linking data on plays and concerts to these resources, a wealth of external data on artefacts and actors enriches our local resources, and we make our local data available in a global context.

Visualizing webs of LCEs over time requires techniques that go beyond the standard features of visualization tools and libraries. The challenge in visualizing such networks is that we have to account for both multimodality and time. In our paper, we explore the possibilities and limitations of visualizations by looking at the networks behind the reception of the French playwright Molière in Dutch theatre in the 17th and 18th centuries. This treatment of the international linkages of local theatrical performances effectively shows how operationalizing cultural life through the concept of, and data on, linked cultural events may assist researchers in 1) mapping cultural life in both quantitative and qualitative ways, and 2) analysing the organisation of cultural life beyond a single event or fixed network of local actors.

²² Dell Hymes (1964), 'Introduction: Toward ethnographies of communication', *American Anthropologist* 66 (6-II), pp. 1-34, p. 13. <https://www.jstor.org/stable/668159>.

²³ <http://www.vondel.humanities.uva.nl/onstage>. Kim Jautze, Frans Blom, Leonor Álvarez Francés (2016) 'Spaans theater in de Amsterdamse Schouwburg (1638-1672). Kwantitatieve en kwalitatieve analyse van de creatieve industrie van het vertalen'. *De Zeventiende Eeuw. Cultuur in de Nederlanden in interdisciplinair perspectief* 32(1), pp. 12-39. DOI: <http://doi.org/10.18352/dze.10000> 6; Kim Jautze, 'ONSTAGE! Presentation at the Conference for "Werkgroep voor de Zeventiende Eeuw" in Nijmegen, 29 August 2015', *EMagazine eHumanities Royal Netherlands Academy of Arts and Sciences* 6, <http://ehumanities.leasepress.com/emagazine-6/recent-events/onstage>.

²⁴ Mascha van Nieuwkerk and Harm Nijboer, 'Nineteenth century concert programs in a digital research environment; the case of Felix Meritis'. Poster presentation at DHBenelux Belval, 9-10 June 2016. http://www.dhbenelux.org/wp-content/uploads/2016/05/60_nieuwkerk_nijboer_FinalAbstract_poster.pdf

1. The Quest for Questions in Digital History: A Comparative View on Werner- and Delors Report on Economic and Monetary Union

Florentina Armaselu and Elena Danescu

1. Introduction

In *The Formation of the Scientific Mind*, Bachelard (2002: 25) considers “the sense of the problem” as the core of the construction of scientific knowledge: “all knowledge is an answer to a question”. Within the framework of Digital Humanities and text analysis, Ramsay (2003: 171, 173) proposes the term of “algorithmic criticism” implying a way to assess, beyond hypotheses validation, “how successful the algorithms were in provoking thought and allowing insight”. In the context of Digital History (Seefeldt and Thomas, 2009) and language study in history (Bertrand et al., 2011), this proposal deals less with how textual analysis confirms/disconfirms previous hypotheses and more with how digital tools help articulate research questions and foster new paths for interpretation. The analysed texts, Werner- and Delors report, were selected for their contrastive, comparative potential and their importance in the Economic and Monetary Union (EMU) history.

2. Thematic snapshot: Werner- and Delors Report

At the Hague Summit (December 1969), an experts committee chaired by Pierre Werner (Prime Minister of Luxembourg) was set up to explore the progress towards EMU in the European Community (EC). The result was the Werner report (1970), which offered a full definition of EMU (3 stages over 1971–80). Goals: achieve irreversible convertibility between the Member States currencies, the complete liberalization of capital movements, the irrevocability of exchange rates, and even a single European currency. Two main principles underpinned this report: gradual realization of EMU and parallelism between economic and monetary convergence. In 1974 the Werner report was suspended. In 1988 was set up a committee charged with the Study of EMU, chaired by Jacques Delors (President of the European Commission). The result was the Delors report (1989) which was appropriating the overall philosophy and structure of the Werner report.

3. Methodology

The quest for questions started as a comparison of the documents, using the corpus analysis framework TXM. TXM has been chosen for its contrastive potential via the specificities feature highlighting what properties are specific, as overuse/deficit (Lafon, 1981), to a part versus the rest of a corpus.

The corpus contains the reports in txt format, as whole and fragments (numbered parts and sections) in separate files. The files were imported into TXM (TXT+CSV) and tagged via TreeTagger (French). Partitions were created for the entire reports and the parts/sections. The analysis was based on:

- lexical table and specificities, nom-adjective query : [frpos="NOM.*"] [frpos="ADJ.*"] (Vmax = 500, Edit = frlemma);
- specificities, part of speech (frpos).

The selection of properties was driven by specificity scores, higher than the TXM default banality threshold (+/-2.0), and relevance. The derived diagrams were used to formulate research questions.

4. TXM Analysis and Questions Formulation

Figure 1 shows a selection of concepts defining the “monetary” aspect of EMU as reflected by the Werner- and Delors report. From the specificities scores/diagram, the following question was formulated:

Q1: How monetary matters (currency, budgetary and fiscal topics) differ in Werner- and Delors report?

Despite of the apparent similarity between the two reports, the diagram shows a contrast within several notions (*inter-Community margins*, *Community currencies* versus *monetary union*).

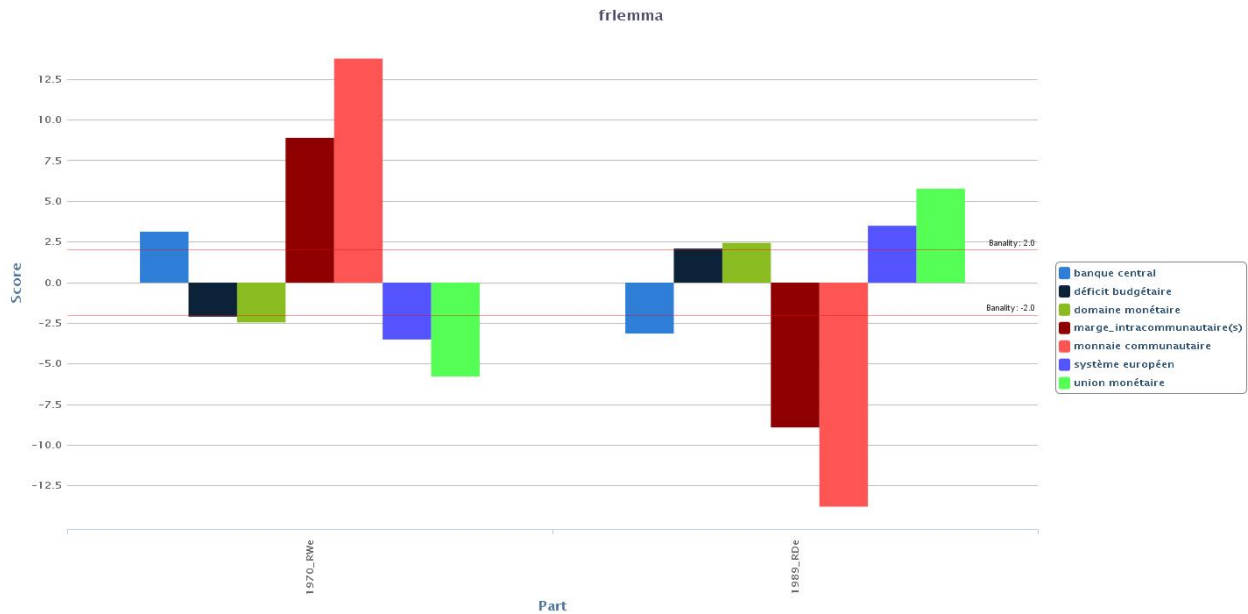


Fig 1. Specificities: EMU “monetary” aspect, Werner- Delors report (**RWe-RDe**) (whole view)

More details on the differences are provided by TXM co-occurrences: “monnaie communautaire” with “fluctuation”, “intervention”, “marge(s)”, focusing on the monetary stabilisation process (RWe); “union monétaire” with “convertibilité totale et irréversible”, emphasising the “monnaie unique” objective (RDe).

Figure 2 illustrates other oppositions related to the “economic” aspect of EMU and the question:

Q2: How the economic matters (economic policy, market) differ in Werner- and Delors report?

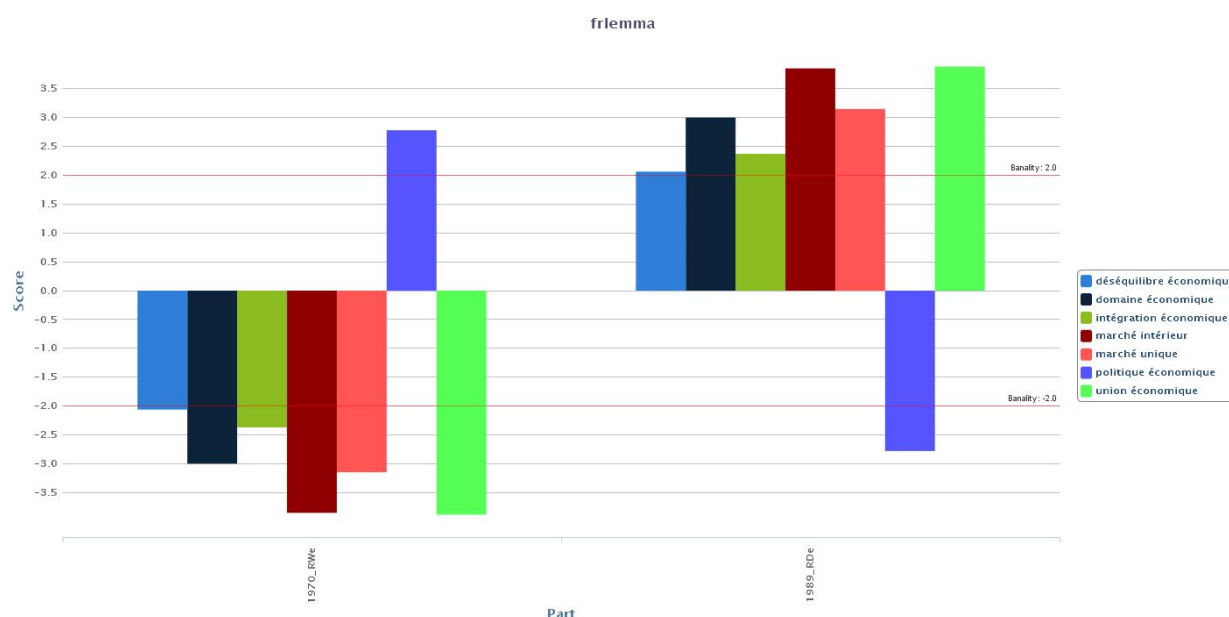


Fig 2. Specificities: EMU “economic” aspect, RWe-RDe (whole view)

The distinction *process/objective* may be further observed via TXM co-occurrences: “politique économique” with “convergence”, “coordination”, “centre de decision” (RWe); “marché intérieur/unique” with “programme d’achèvement”, and “déséquilibre économique” with “corriger” (RDe).

A similar analysis applied to the parts/sections corpus provided further incentives for enquiries on terminological and “actors”-related matters (Q3, Q4).

Q3: Can we speak of an evolution of the EMU terminology between 1970 and 1989 as reflected by the structure of the two documents?

Q4: What influence upon the EMU construction did have the structure of the Werner Committee membership (mainly politicians) and of Delors Committee (mainly central bankers)? What terminology for what people at what moment?

In the European integration process, many concepts evolved from hypotheses to reality between 1970 and 1989. It is why some terms (*central bank, European system, intra-Community margins, monetary union*) are over/under-represented in certain sections. The structure of RWe reflects the distribution of roles – politicians designed the scope, elements and stages of the EMU process (main part); central bankers (appendix 5) set up the technicalities of the European currency and the architecture of the ESCB (*European System of Central Banks*) (Fig. 3, 4).

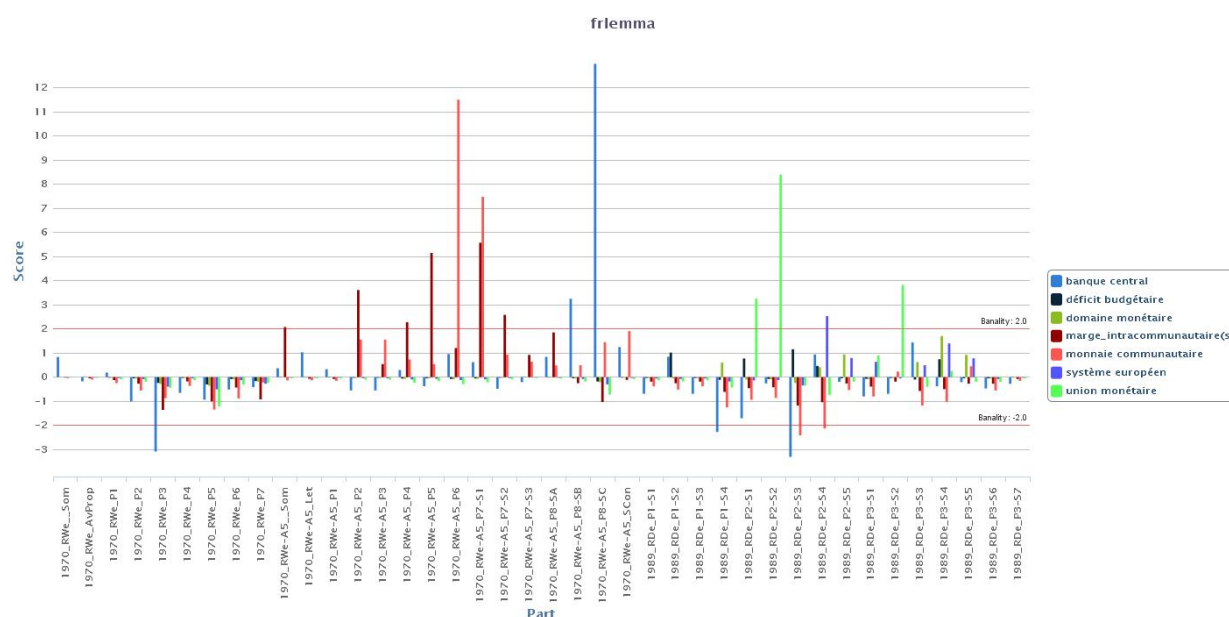


Fig. 3. Specificities: EMU “monetary” aspect, RWe-RDe (structure view)

Q5: The Monetary Union and the Economic Union processes were really designed on a symmetrical and simultaneous basis? A processes granularity analysis.

In RDe, the degree of detail describing the *economic union* process is less than that of the *monetary union*. This may be assumed by looking at the document sections showing high specificity scores for these terms (Fig. 3, 4).

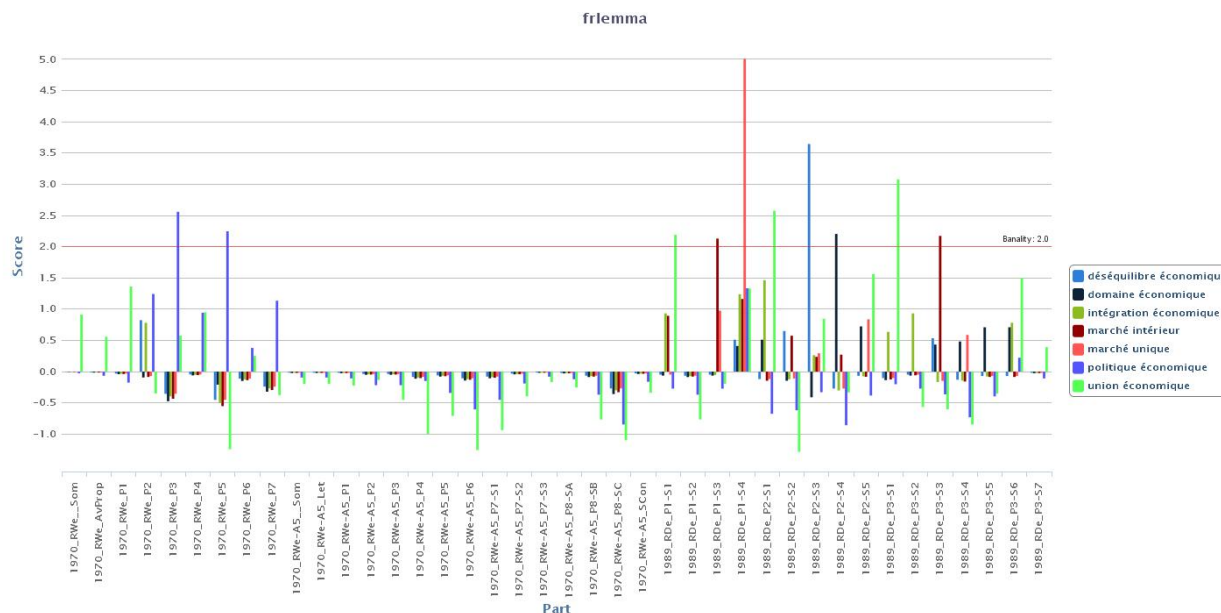


Fig. 4. Specificities: EMU “economic” aspect, RWe-RDe (structure view)

The analysis of specificities computed according to the part of speech (frpos) revealed other salient oppositions related to the use of verbal forms, adjectives, pronouns and citation marks (Fig. 5, 6).

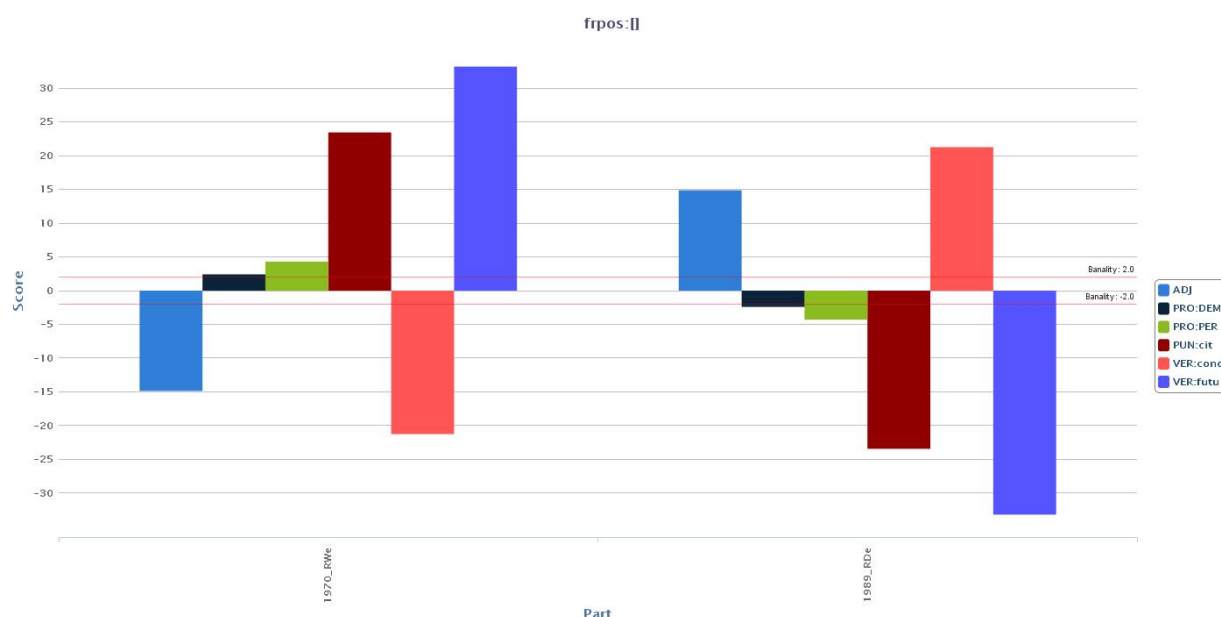


Fig. 5. Specificities: part of speech, RWe-RDe (whole view)

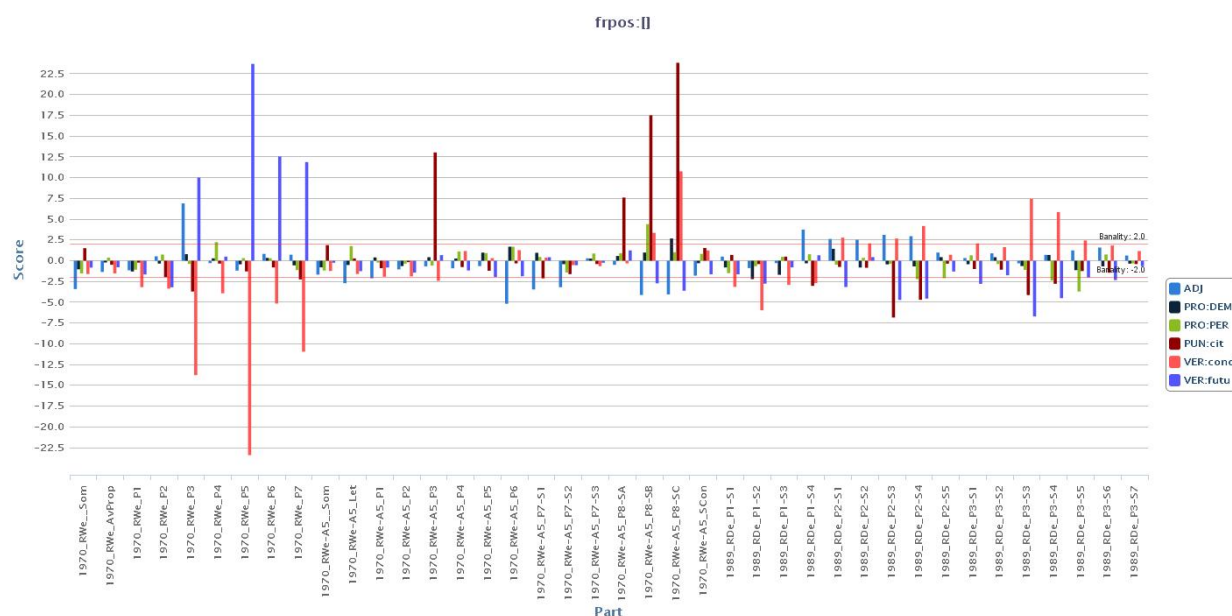


Fig. 6. Specificities: part of speech, RWe-RDe (structure view)

One can observe a dominance of the future verbal form in RWe versus conditionals in RDe, leading to:

Q6: What is behind the range of verbal forms in Werner- and in Delors report? Decoding hidden political meanings and national interests.

Q7: What is the degree of certainty and inter-conditionality between the single market and EMU?

RWe was defining a decade projection for the EMU process while RDe was built upon its first stage achievements but in an uncertain environment. This may elicit further investigation on the verbal forms usage (Q6, Q7).

5. Conclusions

The proposal revisits traditional methodologies in contemporary history and DH from an epistemological perspective: use of comparative textual analysis to formulate research questions.

The first experiments with two crucial EMU documents suggest that digital tools may serve as hypotheses or conclusions validators but also as means of discovering exploration paths in the construction of new knowledge.

References

Bachelard, Gaston. *The Formation of the Scientific Mind: A Contribution to a Psychoanalysis of Objective Knowledge*, Clinamen Press, 2002.

Bertrand, Jean-Marie. Boilley, Pierre. Genet, Jean-Philippe. Schmitt-Pantel, Pauline (éditeurs). *Langue et Histoire*, Paris, Publication de la Sorbonne, 2011.

Lafon, Pierre. 1980. "Sur la variabilité de la fréquence des formes dans un corpus", In *Mots. Saussure, Zipf, Lagado, des méthodes, des calculs, des doutes et le vocabulaire de quelques textes politiques*, N°1, pp. 127-165. http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1008.

Ramsay, Stephen. "Special Section: Reconceiving Text Analysis: Toward an Algorithmic Criticism", *Lit Linguist Computing* (2003) 18 (2): 167-174. DOI: <https://doi.org/10.1093/lc/18.2.167>. Published: 01 June 2003.

Seefeldt, Douglas. Thomas, William G. "Intersections: History and New Media. What Is Digital History?", In *Perspectives on History, The Newsmagazine on the American Historical Association*, May 2009.

TXM, Textométrie, <http://textometrie.ens-lyon.fr/?lang=en>.

Sources

Rapport au Conseil et à la Commission concernant la réalisation par étapes de l'Union économique et monétaire dans la Communauté (rapport Werner). Luxembourg: 8 octobre 1970, document L 6.956/II/70-D. In *Journal officiel des Communautés européennes*, n° C 136, supplément au Bulletin 11/1970, Luxembourg, 11 novembre 1970.

Rapport sur l'Union économique et monétaire dans la Communauté européenne (rapport Delors). 12 avril 1989. In *Europe Documents*. Bruxelles, 20 avril 1989, n° 1550/1551.

2. Transparency as Rupture: Open Data and the Datafied Society of Hong Kong

Rolien Hoyng
Lingnan University

This paper deals with Open Data and the datafication of governance in Hong Kong. It addresses contestations over "transparency" as a techno-political construction that is embodied in, and performed by, the infrastructures and techniques of data-driven governance. Transparency is a site of negotiating distributions of cognition and perception in the context of transformations of citizenship and governance in the datafied society. I specifically inquire into the infrastructures, protocols, techniques, and practices of Open Data, which promises to simultaneously enhance government accountability and stimulate data-driven "smart" governance. Accordingly, I look at

techno-political organizations of digital data and data infrastructures that support particular modes and distributions of cognition and perception (Halpern 2014; Hayles 2014; Kitchin 2014). I distinguish two data regimes respectively revolving around “representation” and “prediction.” I review these issues in relation to the question of globalization. The case study of Hong Kong suggests that datafication does not result in globally homogeneous cybernetic control. Rather, the process of adapting Open Data is (structurally) incomplete, disruptive, and disrupted in the encounter with residual rationalities of statecraft, which means it opens up a field of struggle and contestation. In this paper, “disruption” functions as a methodological device to explore the politics of Open data and datafication at large. Rather than appropriating disruption as a revelatory moment undoing the “black-boxing” of technology per se, my aim is to rethink the politics of transparency and secrecy in more complex terms and inquire into the possibility of activism and intervention (Birchall 2015).

I deploy mixed methods including interviews with actors and analyses of policy documents and technical literature as well as material architectures, formats, protocols, interfaces, and data visualizations. On the basis of examples including the data.gov.hk website and fintech apps, I argue that the two data regimes of “representation” and “prediction” enact particular “fields” of visibility: organized articulations of strategies, techniques, and discourses (Halpern 2014). First, the data regime of “representation” provides cognition and perception in terms of oversight and transparency. Ordering data (capturing, aggregating, and organizing) forms part and parcel of ordering society. Data forms evidence for what exists “out there” and affords referential, descriptive capability. Hence, it is supposed to assist in the production of knowledge and truth. Second, the data regime of prediction, which is afforded by digital data processing techniques and infrastructure, orients perception and cognition onto diagnosis of potential and the prediction of tendencies. Data is generated without a specific question or purpose in mind. Rather than depicting the world, at stake is modeling the world for tactical interventions in shifting patterns and trends (Andrejevic 2013). Distribution of this mode of perception and cognition induces society’s mediation by algorithmic data processing techniques.

Rather than recognizing data regimes in an ideal-typical fashion, my main question addresses the processes of adaptation and the contradictions that emerge due to intersecting of data expediency. This focus underscores Open Data’s paradox of promising fortified transparency and accountability, while simultaneously advancing covert forms of modulation, control, dataveillance, and concentrations of cognition. For instance, citizen-consumers as users of apps are interpellated into positions that seemingly democratize predictive perception and cognition, yet they are simultaneously subjected to dataveillance and algorithmic governance. However, the datafied society does not present itself as a *fait accompli*, in other words, fully operational and all-encompassing. Rather, <mailto:heidichan@LN.edu.hk> adaptation induces instances of (experienced) failure, disruption, and deferment; it generates contradictions, interferences, and articulations between co-existing data regimes and multifarious political rationalities (cf. Chan 2013). These moments might offer possibilities for imagining more radical notions of transparency and secrecy.

If transparency and secrecy are co-constituted, the question is what escapes the particular constructions of transparency in Open Data (Birchall 2015). For instance, if the government opens up certain datasets, does this enable public scrutiny of statecraft or does it merely benefit the expansion of what Easterling (2015) calls extrastatecraft—now by means of data-driven apparatuses belonging to institutions that do not open their own proprietary datasets? How do data and data infrastructures mediate citizens’ relation to private-public governance? To what extent are Open Data activists able to not just reclaim public scrutiny over statecraft but radicalize transparency, for instance by introducing uncontrollable data motility and reversible transitions between data and information? Following a more speculative turn, should transparency always be the goal, or does secrecy have its merits too in order to intervene into the effects of predictive perception and cognition on society?

References

- Andrejevic, Marc. 2013. *Infoglut: How Too Much Information Is Changing the Way We Think*. New York: Routledge.
- Birchall, Clare. 2015. "'Data.gov-in-a-box': Delimiting Transparency." *European Journal of Social Theory*. 18(2):
- Bratton, Benjamin. 2015. *The Stack: On Software and Sovereignty*. Cambridge MA: MIT Press.
- Chan, Anita Say. 2013. *Networking Peripheries: Technological Futures and the Myth of Digital Universalism*. Cambridge MA: MIT Press.
- Easterling, Keller. 2015. *Extrastatecraft: The Power of Infrastructure Space*. London: Verso.
- Halpern, Orit. 2014. *Beautiful Data: A History of Vision and Reason since 1945*. Durham: Duke University Press.
- Hayles, N. Katherine. 2014. "Cognition Everywhere: The Rise of the Cognitive Nonconscious and the Costs of Consciousness." *New Literary History* 45 (2): 199-220.
- Kitchin, Rob. 2014. *The Data Revolution Big Data, Open Data, Data Infrastructures and Their Consequences*. Los Angeles: Sage.
- Ong, Aiwha. 2006. *Neoliberalism as Exception*. Durham: Duke University Press
-

3. Oral history online – User perspectives and behavior in a transforming WW2 memory culture

Dr. Susan Hogervorst | Open Universiteit Nederland/Erasmus University Rotterdam | susan.hogervorst@ou.nl

Since the 1980s, Second World War (WW2) memory culture has been increasingly characterized by the forthcoming disappearance of the eyewitness generations. One way in which this problem has been addressed, is by recording eyewitness testimonies. By now, multiple oral history collections have been created throughout the western world, in which ten thousands of interviews have been preserved on audio and video (Apostolous and Pagenstecher 2013, Keilbach 2013). Currently, we see a shift from collecting and preserving testimonies to disclosing them for wider audiences (Scagliola and F. de Jong 2014; S. de Jong 2013; Bothe and Lücke, 2013). This is partly due to technological developments, but also to the dynamics of WW2 memory culture, of which transmitting eyewitness memories onto younger generations has become a key feature (Wieviorka 2006; Erll and Rigney 2009; Hogervorst 2010; Sabrow and Frei 2012). How will the disappearance of the eyewitness generations affect the transmission of eyewitness memories, and what role do online interview collections play in this process?

Methods and data

The aim of my postdoc research project is to provide substantiated data on this matter, partly acquired by analyzing content, use, and users of an online video interview collection: the Dutch web portal 'Getuigenverhalen.nl'. This portal gives access to circa 500 video interviews with eyewitnesses about different WW2 related topics. I deploy this interview collection as a digital barometer of current, transforming memory culture: through the web statistics, an online questionnaire, focus group interviews with (student) history teachers, and screen recordings of their interaction with the

website while selecting interview fragments they would use in a lesson about WW2. This innovative approach not only underlines the value of incorporating digital sources and methods into the field. It also enables foregrounding the user and his/her agency in the analysis of the rather top-down, officially supported process of WW2 memory transmission, while users often remain elusive in traditional memory studies (and historical) research (Kansteiner 2002; Erll and Rigney 2009).

Findings and discussion

The findings indicate both continuity and change regarding the use of WW2 video interviews compared to live testimonies in classrooms. The portal is first and foremost an online archive; it is not explicitly meant as an educational tool. Inquiries in the educational field in the Netherlands point at an interest in, but also at an unfamiliarity with such collections and their didactical possibilities. This is confirmed by both the questionnaire and the web statistics. Only few respondents identify themselves as teacher or student, and relatively many do not find what they were looking for. Site search is not used often, and the mostly watched interviews are the ones highlighted on the homepage. Two focus group interviews with an international group of student history teachers offered an opportunity to get a closer and more in-depth view on the portal's use, and – to a scholar of cultural memory more importantly – on users' selection criteria of relevant material out of the quite abundant reservoir of eyewitness testimonies available. First, according to the participants, a suitable interview fragment should 'bring the past closer' (which was to be achieved in different manners). Indeed, the participants could quite easily find fragments that suited these purposes. Second, suitable fragments should (according to the participants) confirm existing historical knowledge. The latter indicates a more fundamental, both ethical and epistemological view on the position and value of eyewitnesses and their functioning as sources of historical knowledge.

Both aspects correspond to the way participants would use live testimonies in their lessons. The plenary evaluation of the selected fragments and the criteria used, pointed at an important difference: the distance through the screen. This distance enabled raising critical questions about the nature and value (reliability) of oral testimonies, which is rather uncommon in settings in which eyewitnesses are physically present. Another characteristic of searchable interview collections, that they enable comparing different testimonies and experiences, and therewith supplement or challenge (besides confirm and illustrate) existing historical knowledge and perspectives, was not mentioned by the participants. This might point at the fact that working with digital testimonies is still in an early stage in the Netherlands. Currently, in Germany, Austria, and the United States, educational projects are developed around WW2 video testimonies. Besides recommendations for improvements of the Dutch portal website, this study contributes to a more critical, and didactically more relevant interaction with testimonies in Dutch history education, as well as to a better understanding of current transforming memory culture.

References

- Apostolous, N. and C. Pagenstecher (eds.), *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt* (Berlin 2013).
- Bothe, A. and M. Lücke, 'Shoah und historisches Lernen mit virtuellen Zeugnissen', P. Gautschi et al. (eds.), *Shoah und Schule. Lehren und Lernen im 21. Jahrhundert* (Zürich 2013) 55-74.
- Erll, A. and A. Rigney, *Mediation, remediation, and the dynamics of cultural memory* (Berlin/New York 2009).
- Hirsch, M. and L. Spitzer, 'The witness in the archive. Holocaust studies/memory studies', *Memory Studies* vol. 2 (2009)2, 151-170
- Hogervorst, S., *Onwrikbare herinnering. Herinneringsculturen van Ravensbrück in Europa, 1945-2010* (Hilversum 2010).

Huijgen, T., & Holthuis, P. (2016). Dutch voices: exploring the role of oral history in Dutch secondary history teaching. In D. Trškan (Ed.), *Oral history education: dialogue with the past*. (1 ed., Vol. 1, pp. 43-58). Ljubljana: Slovenian National Commission for UNESCO.

Jong, S. de, 'Im Spiegel der Geschichten: Objekte und Zeitzeugenvideos in Museen des Holocaust und des Zweiten Weltkrieges', *Werkstatt Geschichte* 62 (2013) 19-41.

Kansteiner, W., 'Finding meaning in memory. A methodological critique of collective memory studies', *History and theory* 41 (2002) 179-197.

Keilbach, J., 'Collecting, Indexing and Digitizing Survivors. Holocaust Testimonies in the Digital Age', A. Bangert et al. (eds.), *Holocaust Intersections. Genocide and Visual Culture at the New Millennium* (London 2013) 46-63.

Scagliola, S. and F. de Jong, 'Clio's talkative daughter goes digital', R. Bod et al. (eds.), *The Making of the Humanities, Volume III: The Modern Humanities* (Amsterdam 2013) 511-526.

Sabrow, M. and N. Frei (eds.), *Die Geburt des Zeitzeugen nach 1945* (Göttingen 2012).

Wieviorka, A., *The era of the witness* (Ithaca 2006).

1. Collections as networks, Uncovering information exchanges and information networks in the collections of the Meertens Institute (KNAW)

Douwe A. Zeldenrust
Meertens Institute (KNAW)

This paper is about uncovering information exchanges and information networks in humanities research collections. Most humanities researchers focus on obtaining data from research collections, without realizing that those collections also can be seen as the results of epistemological experiments. That is: every collection is the outcome of the process of gathering information and therefore interconnected with the presuppositions, foundations and the activities that have led to the knowledge it contains. Charles Jeurgens (2012) states about this connection that: ‘(...) understanding that bond has to precede understanding the records’ (p. 51). The ‘bond’ Jeurgens writes about, is not only dictated by the goal(s) forming the collection, but is also determined by (among other things) the cultural, administrative, scientific and social climate. Moreover, it is dependent on the individuals who were collecting, their scientific experience, their interests and personalities.

The paper will reflect on the issues of extracting, visualising and processing this context information, using the concept of ‘deep networks’. Charles van de Heuvel introduced this concept in his article ‘Mapping knowledge exchange in early modern Europe intellectual and technological geographies and network representation’ (2015). The concept allows the contextualisation of networks and the visualisation of uncertainty while creating layers of historical sources in multiple perspectives. Furthermore, it combines pattern recognition in textual and visual big data with traditional hermeneutic methods. The vast collections of the Meertens Institute (Royal Netherlands Academy of Arts and Sciences) will be used as a use case in order to make the first steps in realizing these networks within the framework of archival studies (Meertens, 2016).

The collections of the Meertens Institute have been accumulated in a period of over 80 years and concentrate on the diversity in language and culture in the Netherlands (Jongenburger, 2013). Access to the more than 15 terabytes of data, 6000 hours of (digital) audio and 2 kilometres of archival material is provided by a record keeping system containing data about, amongst other things, the researchers involved in collecting. The information captured in this record keeping system creates the first layer of the network. A second layer of information, regarding the provenance of the collections, is extracted from the annual reports of the Meertens Institute. Those reports contain information about, for instance, the acquisition of the collections. And a third layer of information is obtained from the Biographical Portal of the Netherlands (Biografischportaal, 2016). This online reference work contains short descriptions of the lives of persons (amongst them prominent scholars and influential managers of research institutes) who distinguished themselves or played a role of some significance in the past in the Netherlands.

The objective of this research is threefold: first it would demonstrate that building such a network is feasible. The web-based software platform Palladio will be used for processing the data and visualizing the network (Palladio, 2016). As this research is ongoing, experiments with other, more versatile network analysis tools, such as Nodegoat, will be considered (Nodegoat, 2016).²⁵ The

25 Various data visualization platforms and network architectures have been developed. For visualizing data Palladio is one of the platforms that is advised as a network visualization tool for the humanities (Düring,

network will consist of hundreds of nodes and thousands of (potential) edges in order to include the relations among the most prominent persons involved in collecting the information. Second, this method can, with local modifications, be reused by other humanities researchers to generate networks for archival studies. And third, the outcomes will be incorporated in my PhD research, which is about the history of the collections of the Meertens Institute. As this PhD research started in January 2016 and is ongoing, this paper will show the first results.

References:

Düring, M. (2016). On Dilettantes and Dialogues in Digital History. *European Social Science History Conference 2016*.

Heuvel, C. van den (2015). Mapping Knowledge Exchange in Early Modern Europe. *International Journal of Humanities and Arts Computing*, 9 (1), 95-114.

Jeurgens, K.J.P.F.M. (2012). Information on the move. Colonial archives: pillars of past global information exchange. *Colonial Legacy in South East Asia. The Dutch Archives*. Eds. K.J.P.F.M Jeurgens, A.C.M Kappelhof & M. Karabinos. 's-Gravenhage: Stichting Archiefpublicaties. 45-65.

Jongenburger, W, A.W.H. Jansen & D.A. Zeldenrust (2013). *Collectieplan Meertens Instituut, 2013-2018*. Amsterdam: Meertens Instituut.

Websites:

<http://ckcc.huygens.knaw.nl> (Accessed December 08, 2016)

<https://nodegoat.net> (Accessed December 08 17, 2016)

<http://palladio.designhumanities.org> (Accessed December 08, 2016)

<http://www.biografischportaal.nl> (Accessed December 08, 2016)

<http://www.meertens.knaw.nl> (Accessed December 08, 2016)

2. Mapping Controversies of Digital Curation

Dana Mustata

University of Groningen, NL

The emergence of digital technologies and digitized data in humanities research – a phenomenon that has been shaping the contours and the incentives behind the organization of digital humanities as field of study – has raised more questions than has answered any. Are digital technologies changing our research practices and if so, how? Do they incite new research questions? Are established fields of study in the humanities drastically altering in the face of these new phenomena that have technology at their centre? What is new and what is old in the way we do research in digital environments? If there is any underlying assumption traversing all these questions is that digital humanities is a ‘transformative practice’ (Svensson, 2009). It has been primarily the difficulty to describe and explain what it is that is changing in our research practices. This paper tackles this particular concern.

Has there been a shift in our traditional research practices rooted in the analogue era? What does this shift consist of? How do we redefine ourselves from traditionally analogue researchers into

2016). Nodegoat was used in the project ‘Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic’ (Huygens, 2016).

digital humanities researchers? These questions contribute to furthering the definition of digital humanities as a field of study and more specifically, to redefining the practice of doing scholarship with digital technologies and digitized data. Without advocating for essentialist, stable and fixed definitions of digital humanities as an arena in which scholarship is produced, I am particularly interested in what Bruno Latour (1991) calls the ‘socio-logics’ characterizing digital humanities as a transformative field of study. Socio-logics refer to how knowledge is mobilised, constructed and accumulated in the face of a ‘controversy’.

“The word “controversy” refers here to every bit of science and technology which is not yet stabilized, closed or “black boxed” ... we use it as a general term to describe *shared uncertainty*. (Macospol, 2007: 6, cited in Venturini, 2010: 260).

In other words, as Tommaso Venturini explains:

“controversies are situations where actors disagree (or better, agree on their disagreement). The notion of disagreement is to be taken in the widest sense: controversies begin when actors discover that they cannot ignore each other and controversies end when actors manage to work out a solid compromise to live together. Anything between these two extremes can be called a controversy.” (Venturini, 2010: 261).

The starting point of my paper is thus, approaching digital humanities as a controversial arena, one in which researchers, tools, tool developers and data providers – humans and non-humans, actors and actants in Latour’s terms – collide; an arena in which scientific and technological clashes play out. It is through the chains of association between researchers, tools, tool developers and data providers as well as through the transformations prompted by the clashes between these actors and actants that the socio-logic of this new field of study is rendered visible.

The paper will ‘map the controversies’ of curating digital objects in a virtual research environment. These controversies relate to translating academic knowledge into tools design and implementation, translating historical narratives into user functionalities, finding a shared work language and collaborating at the intersection of different fields of expertise and fields of knowledge production.

Mapping controversies is a hands-on method rooted in the ANT tradition of thought, which explores, describes, visualizes and makes sense of issues that emerge at the intersection of collaborative work done between – in this particular case - researchers, tools, designers, tool developers and digital data providers. This particular method provides insights into practices of working with digital tools and digitised data and the subsequent process of knowledge production.

The paper will map controversies through the practices of designing, researching and curating the virtual exhibitions (VEs) on the online platform www.euscreen.eu. This is a platform that makes freely accessible thousands of audiovisual items originating from 21 content providers in Europe. The VEs were curated by researchers in collaboration with 1) content providers (CPs) consisting of audiovisual archives and responsible for co-selecting and uploading their content to the ‘Special Collections’ on platform; 2) tool developers in charge of developing the VE builder and all the user functionalities around it; 3) designers responsible for the design of the frontend of the exhibition, the design of the user experience as well as for mediating the common grounds between the researcher’s needs and the potentials of tool development; 4) the VE builder which allowed the researchers to curate their exhibitions. The researchers drafted the content selection strategy for the CPs; viewed, researched, further selected and then bookmarked the content uploaded to the Special Collections; defined and advised on the development and design of the VE builder through wireframing, paper prototyping and joint work sessions with the designers and tool developers; and last but not least, curated their virtual exhibitions as an end result of all these collaborative work practices.

Taking a practice-oriented anthropological approach to the researchers' journey through curating the VEs, the paper will explore digital curation through what Latour (1991) called the 'chain of associations' and the 'series of transformations' underwent by the actors and actants involved as well as through the 'translations' that took place throughout the collaborative work process, which saw the initial enunciations of the VEs turn into the final products that were published online.

By mapping the associations that researchers entered into, the transformations they underwent as part of these associations and the translations that took place from their first VE ideas to the final curated objects online, this paper tries to pin down what it is that changes in the practice of doing humanities research when knowledge is (co)produced with digital tools, digitized data and at the intersection at different fields of expertise.

Making sense of digital humanities as a transformative practice of (co-)producing knowledge is a fertile ground to come to terms with this emerging field of study. It helps us understand 'digital practices' in terms of what humanities scholars do with digital tools in digital environments, to paraphrase Couldry's (2010) understanding of 'media practices'. I argue, thus that understanding the production of knowledge in digital humanities becomes an archaeological act of retracing associations and transformations through different spaces of expertise, different actors and actants, lending itself to what Foucault called 'principles of discontinuity'. This may help bridge the gaps between digital technologies and (analogue) researchers, technicians and scientists that are at the core of controversies in the field.

References

Nick Couldry, 'Theorising Media as Practice' in *Social Semiotics*, Vol. 14, Issue 2, 2004, pp. 115-132. Published online: 13 Oct 2010, <http://dx.doi.org/10.1080/1035033042000238295>

Michel Foucault, *Archaeology of Knowledge*, Routledge, 1972

Bruno Latour, 'Technology is Society Made Durable' in: J. Law, ed., *A Sociology of Monsters Essays on Power, Technology and Domination*, Sociological Review Monograph N°38, pp. 103-132, 1991

Patrik Svensson, 'Humanities Compuring as Digital Humanities', *Digital Humanities Quaterly*, 3 (3), 2009

Tommaso Venturini, 'Diving in magma: how to explore controversies with actor-network theory' in: *Public Understanding of Science*, 19 (3), 2010, pp. 258-273

3. Research opportunities for the archived web in the Benelux

Sally Chambers, Ghent Centre for Digital Humanities, Ghent University

Peter Mechant, Media, Innovation and Communication Technologies (MICT), Ghent University

Kees Teszelszky, National Library of the Netherlands

Yves Maurer, National Library of Luxembourg

Web-archiving or collecting portions of the web to ensure the information is preserved in an archive, began in 1996 with the Internet Archive initiative²⁶ and its well-known digital archive 'The Wayback Machine'²⁷. Others have followed, from national and state libraries and archives to museums and

²⁶ <https://archive.org/>

²⁷ <https://archive.org/web/>

nonprofits such as ‘Common Crawl’ which corpus contains more than 3.14 billion web pages and about 250 TB of uncompressed content²⁸.

Although most researchers in the humanities still need to begin to explore the potential of these archives, some projects have already investigated their potential, e.g. the BUDDAH project²⁹ (Big UK Domain Data for the Arts and Humanities) which awarded bursaries to researchers to carry out research in their subject area using the UK web archive, or the RESAW network³⁰ (Research Infrastructure for the Study of Archived Web Materials) which aims at promoting a collaborative European research infrastructure for the study of archived web materials. Despite such initiatives, researchers in the humanities still struggle with the computational turn of their field on theoretical, methodological (e.g. develop theoretical and methodological frameworks within which to study web archives) and practical levels (e.g. they lack expertise and knowledge to use web archives and to apply digital methods or big data techniques on their corpus).

Although geographically very close (the history of) national web archiving is very different for the three Benelux countries:

In the Netherlands, the National Library already started in 1992 with mapping the Dutch web by compiling web directories or web lists. A first web archiving pilot was conducted in 2003 and web-archiving started as a regular activity in 2007 using a selective harvesting strategy based on a selection of the existing web directory (governmental, cultural and academic websites, sites that mirror trends on the web, and ‘endangered’ websites which are considered as Dutch digital cultural heritage). As per January 2017, 12,000 websites have been harvested with Heritrix³¹ (26 TB of compressed .arc files, 211 million URLs). A linguistic analysis of the collection has not been done yet, but 368 Frisian websites are included. The Dutch web archive is available in the reading rooms of the National Library of the Netherlands or researchers can request access to the data for specific projects³².

In Luxembourg, a pilot project for web-archiving was undertaken in 2005 and subsequently the legal deposit law was extended in 2009 to also cover content published on the web. Due to funding issues, the regular harvests for the .lu domain and other websites hosted in Luxembourg only started in August 2016. A second crawl finished in January 2017. These crawls were supplemented with data from a number of targeted crawls of governmental sites. Currently, the archive contains 15 TB of compressed warc files (250 million URLs) with around 40% of the hosts in the ccTLD .lu and 35% in the .com. Similar to the Dutch archive, a detailed linguistic examination of the Luxembourg collection has not been done yet, but a basic linguistic analysis shows the presence of 30% English, 30% French, 15% German, 5% Luxembourgish and a ‘long tail’ of other languages. The Luxembourg web archive will be available in the reading rooms of the national library and researchers can be granted access to the underlying dataset on case-by-case basis.

Although the .be domain was introduced in June 1983³³, the Belgian web is currently not systematically archived. As of February 2017, 1.561.932 domains are registered by DNS Belgium³⁴. Without a Belgian web archive, the content of these websites will not be preserved for future generations and a significant portion of Belgian history will be lost forever. In December 2016, a pilot

²⁸ <http://commoncrawl.org/>

²⁹ <http://buddah.projects.history.ac.uk/>

³⁰ <http://resaw.eu/>

³¹ <https://webarchive.jira.com/wiki/display/Heritrix>

³² <https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving>

³³ History of the Belgian web: <https://www.dnsbelgium.be/en/history>

³⁴ DNS Belgium: <https://www.dnsbelgium.be/en>

web-archiving project called PROMISE (PReserving Online Multiple Information: towards a Belgian StratEgy) was funded. The aim of the project is to (i) identify current best practices in web-archiving and apply them to the Belgian context, (ii) pilot Belgian web-archiving, (iii) pilot access (and use) of the pilot Belgian web archive for scientific research, and (iv) make recommendations for a sustainable web-archiving service for Belgium. This pilot project is considered as a first step towards implementing a long-term web archiving strategy for Belgium. Similar to Luxembourg, Belgium has various official languages³⁵ that will need to be considered during the pilot phase.

From a web-archivist perspective, a key challenge is how to collaborate on joint web-archiving initiatives to enable ‘trans-national’ research opportunities, for example, by taking either a site-, topic-, or domain-centric archiving approach, or by unifying methodological approaches for discovery, acquisition and description of web content. From the viewpoint of a researcher in the humanities, web-archives are rich ‘born-digital’ resources, which can analysed alongside other digitised and analogue sources in a wide range of humanities subject areas.

For researching the archived web in the Benelux, possible ‘tri-national’ research questions could include a linguistic analysis of the ‘Benelux web’, or a geo-spatial analysis of the geographic distribution of web-domains across Benelux region³⁶. Similarly, research questions could focus on just two of the Benelux countries, such as the geographic distribution of Dutch-language websites in the Netherlands and Flanders; or the German-language websites in Belgium and Luxembourg. Furthermore, as many European Union institutions (with websites in the .eu domain) are located within the Benelux region, this also offers a further wealth of opportunities for humanities researchers.

While the increased availability of such (big) born-digital datasets opens up the opportunities for using computational research methods, it also points to the need to uptake new skills. It will be therefore be important to establish a range of standard tools and methods which are widely accepted for archived web research³⁷. Despite these challenges, the archived web offers substantial opportunities for digital humanities researchers, both in the Benelux and beyond.

This paper discusses a) the potential of web archives for digital humanities researchers, b) introduces the web-archives in The Netherlands, Luxembourg and Belgium and c) presents the possibilities for trans-national research that collaboration between the Benelux web-archives could enable.

³⁵ The three official languages of Belgium are Dutch, French and German with English also being widely used. Official languages of Belgium: https://en.wikipedia.org/wiki/Languages_of_Belgium

³⁶ DNS Belgium has mapped the geographic distribution of web-domains by local authority across Belgium, see: <https://www.dnsbelgium.be/whois/stats>. The usefulness of extending this mapping could be extended to the whole Benelux region.

³⁷ For examples, see: Truman, Gail. 2016. Web Archiving Environmental Scan. Harvard Library Report: <https://dash.harvard.edu/handle/1/25658314>

Be FAIR or be square: Stakeholders' perspectives on data quality in the Digital Humanities

Reinier de Valk, Data Archiving and Networked Services (DANS)

Vanessa Hanneschläger, ÖAW-ACDH (Austrian Centre for Digital Humanities)

Klaus Illmayer, ÖAW-ACDH (Austrian Centre for Digital Humanities)

Francesca Morselli, Data Archiving and Networked Services (DANS)

Emily Thomas, Data Archiving and Networked Services (DANS)

Introduction

Digital data is created every day. Not only have cultural and research institutes been massively digitising their analogue content over the past decades (*digitised* objects), but research institutes and individual researchers are also constantly producing new digital data (*born-digital* objects). This is not a new revelation: within the natural sciences, researchers have been using and producing structured (and, more recently, machine-readable) data for centuries. However, over the last decades the research landscape has been changing: within the social sciences and humanities (SSH) disciplines, too, the use of existing digital data and the production of new digital data has increased enormously [2, 12, 13, 16]. This entails several issues that must be addressed [5, 9, 18, 23].

The necessity to preserve and ensure reusability for the increasing quantity of this data, which tends to be quite heterogeneous, has made the issue of *data quality* a specifically urgent one [1, 15, 17]. In order to deposit research data in a trusted repository, it needs to meet a minimum set of quality criteria, such as completeness, reliability, and correct formal structure by means of the implementation of interoperable or discipline-specific standards [3, 22].

Moreover, new actors as well as new relationships among them have emerged -- a consequence of the reuse and sharing of research data among researchers and institutions [20]. Not only researchers and research institutions, but also cultural heritage institutions, research infrastructures and European projects -- all of which can be referred to as *stakeholders* -- are now heavily involved in data exchange processes, with the aim of increasing data interoperability and visibility.

Against such a complex background, however, it is difficult to develop and mutually agree on a truly shared vision of what high-quality data is, and what is required to achieve it. One innovative approach to reach common ground is applying the FAIR principles.

The FAIR principles

Following a life sciences workshop in Leiden entitled *Jointly designing a data FAIRPORT* in 2014 [4], a minimal set of community-agreed guiding principles were formulated by a diverse group of stakeholders, sharing an interest in scientific data publication and reuse. This was in order to make data more easily discoverable, accessible, appropriately integrated and reusable, and adequately citable for both machines and people. The principles that were constructed here are now well known as the FAIR principles [6, 7, 8, 24], and act as a guide to data publishers and stewards rather than being a standard or specification. Although these principles were conceived within a life sciences context, social sciences and humanities also face similar issues as they become more digitised, making the topic of FAIR data management also applicable to these fields. In simpler words, the FAIR principles provide a set of mileposts for data producers and publishers to help ensure that all data will be *Findable* (defined by a persistent identifier and detailed metadata), *Accessible* (well-defined license and access conditions), *Interoperable* (ready to be combined with other data by humans and

machines: standardised formats and vocabulary) and *Reusable* (ready to be reused in future research and processed using computational methods).

Stakeholders

When it comes to repository and data quality, the main factor shaping individual needs and requirements is not the discipline the data comes from, but rather the *type* of stakeholder, which is crucial for the perspective on the data and the necessities that come with it. In order to motivate stakeholders to commit to the FAIR principles, the different types first have to be identified and their specific interests have to be investigated. Examples of stakeholders are research communities, funders, data archives, research infrastructures, projects, and cultural heritage institutions. It is important that these groups are brought together to align their different perspectives on data as producers, consumers and providers. For instance, *findability* from a broader perspective might also mean having a user-friendly interface for a researcher to find datasets, whilst for a research infrastructure, the availability of substantial metadata would be the core interest. Therefore, different strategies of communicating and implementing the FAIR principles will be necessary to reach the various types of stakeholders.

Moreover, bringing together different stakeholders allows discussions for collaborations in the implementation of the FAIR principles and sharing experiences of ongoing efforts. Especially when it comes to data quality and data exchange, a discussion of the general framework of the FAIR principles can help to better coordinate the different approaches and interests of stakeholders.

The proposed panel

We propose a panel discussion with representatives of different stakeholders, both current and potential future FAIR implementers. Their discussion will focus on the application of the FAIR principles to improve data quality, formulating FAIR data management requirements (e.g., by funders) and assessing the quality of datasets (e.g., by repositories). This will help determine common approaches as well as variations in perspective. The focus will be on the exchange between those who already (started to) implement the FAIR principles and those who have not yet done so; this will help analyse which goals are attainable by the various stakeholders. We plan to invite representatives of the following types of stakeholders:

- researchers
- research institutes
- cultural heritage institutions
- research infrastructures
- projects
- funders.

The following (deliberately slightly controversial) seven statements are intended to guide the discussion:

1. Data quality is compromised by changing research methods and increased collaboration among stakeholders.
2. As a consequence, stakeholders do not sufficiently address the challenge of guaranteeing data quality due to changing research methods and techniques.
3. Data quality (including implementations of FAIR) is not high enough on the agenda of the various stakeholder groups.
4. Therefore, stakeholders should implement the FAIR principles, even if this means that current approaches have to be adapted.
5. Stakeholders should raise awareness about using the FAIR principles.
6. Data producers are responsible for ensuring the FAIRness of their data.

7. The implementation of FAIR principles should be monitored in institutions and/or among different stakeholders.

Panelists

(to be confirmed shortly)

Selected bibliography

- [1] Batini, C., and Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques*. Berlin: Springer.
- [2] Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.
- [3] Brown, A. (2008). *Selecting File Formats for Long-Term Preservation*. <https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf>
- [4] Data FAIRport. *Data FAIRport conference: Jointly designing a data FAIRport*. <http://www.datafairport.org/component/content/article/8-news/9-item1>
- [5] Dix, A., Cowgill, R., Bashford, C., McVeigh, S., and Ridgewell, R. (2014). Authority and judgement in the digital archive. In 1st Digital Libraries for Musicology Workshop, London, UK.
- [6] Dutch Techcentre for Life Sciences. *FAIR Data*. <http://www.dtls.nl/fair-data/>
- [7] Dutch Techcentre for Life Sciences. *GO-FAIR initiative*. <http://www.dtls.nl/go-fair/>
- [8] Force11. *Guiding principles for Findable, Accessible, Interoperable and Re-usable data publishing version b1.0*. <https://www.force11.org/fairprinciples>
- [9] Giaretta, D. (2011). *Advanced digital preservation*. Berlin: Springer.
- [10] Griffin, G., and Hayler, M., eds. (2016). *Research methods for reading digital data in the Digital Humanities*. Edinburgh: Edinburgh University Press.
- [11] Hayler, M., and Griffin, G. eds. (2016). *Research methods for creating and curating data in the Digital Humanities*. Edinburgh: Edinburgh University Press.
- [12] Kaplan, F. (2015). A map for big data research in digital humanities. *Frontiers in Digital Humanities* 2(1): 1-7.
- [13] Lane, R. J. (2016). *The big humanities: Digital Humanities/digital laboratories*. London: Routledge.
- [14] LERU (2013). *LERU roadmap for research data*. http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf
- [15] NISO (2007). *A framework of guidance for building good digital collections. 3rd ed.* <http://www.niso.org/publications/rp/framework3.pdf>
- [16] Owens, T. (2011). Defining data for humanists: Text, artifact, information or evidence? *Journal of Digital Humanities* 1(1): n.p. <http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/>
- [17] Peer, L., Green, A., and Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation* 9(1): 263-291.
- [18] Pryor, G., ed. (2012). *Managing research data*. London: Facet Publishing.

- [19] Purdy, J. P., and Walker, J. R. (2010). Valuing digital scholarship: Exploring the changing realities of intellectual work. *Profession 1*: 177-195.
- [20] Quan-Haase, A., Suarez, J. L., and Brown, D. M. (2014). Collaborating, connecting, and clustering in the humanities: A case study of networked scholarship in an interdisciplinary, dispersed team. *American Behavioral Scientist* 59(5): 565-581.
- [21] Terras, M. (2010). Digital curiosities: Resource creation via amateur digitization. *Literary and Linguistic Computing* 25(4): 425-438.
- [22] Tjalsma, H., and Rombouts, J. (2011). Selection of research data: Guidelines for appraising and selecting research data. The Hague: DANS.
- [23] Van Zundert, J. (2012). If you build it, will we come? Large scale digital infrastructures as a dead end for Digital Humanities. *Historical Social Research* 37(3): 165-186.
- [24] Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(160018): 1-9.

A Pragmatic Approach to Understanding and Utilizing Events in Cultural Heritage

Lora Aroyo¹, Marnix van Berchum⁴, Lizzy Jongma³, Willem Robert van Hage⁵, Gerard Kuys⁶, Susan Legene¹, Annelies Van Nispen³, Jacco van Ossenbruggen², Lodewijk Petram⁴ and Piek Vossen¹

¹ Vrije Universiteit Amsterdam, The Netherlands

² Centrum Wiskunde & Informatica (CWI), The Netherlands

³ NIOD Instituut voor Oorlogs-, Genocide- en Holocauststudies, The Netherlands

⁴ Huygens ING, The Netherlands

⁵ Netherlands eScience Center, The Netherlands

⁶ Nationaal Archief, The Netherlands

Introduction

Cultural heritage institutions are continuously rethinking the access to their collections to allow the public as well as scholars and professionals to interpret and contribute to their collections. Their collections are challenged with advancement of the Web. They need to be presented in a sustainable way online, and to be instantly searchable and understandable for experts and lay audiences [1]. Hermeneutics is humanities theory of interpretation. Currently it is amended to *digital hermeneutics* to form the appropriate context to think about providing access to and interpretation of online cultural heritage collections [2].

Important role in the *interpretation of cultural heritage collections* play ‘*historic events*’, which meaning keeps being re-discovered and re-interpreted in light of modern discussions. History changes over time and with the presence of the social Web it is under continuous evolvement. “It is not only ‘grand’ historical events that are subject to changes in interpretation. Single words, concepts, ideas and books can also have different meanings across time, space and social groups.” [3]. Automatic text analysis techniques provide the means to mine large amounts of unstructured data and give scholars access to ‘big data’. To understand better this ‘big data’ we observe a shift towards deeper data mining focussed on the retrieval of meaningful units, e.g. *answers, entities, events, discussions, and perspectives*. Additionally, we also observe, a push towards the automatic creation of knowledge graphs that are populated with rich semantic units, e.g. *entities, relations, activities, events* provide possibilities of diving into more the details and address more complex questions. All this comes as a response to the need to understand better ‘events’ and their semantic structure and thus help, on the one hand, heritage institutions *assigning meaning and value to online collection objects*, and on the other hand, help humanities scholars in the *exploration and contextualization of their tasks* [3].

Methodology

This work is motivated by the (1) demands for facilitating deeper understanding of online cultural heritage collections, and by the fact that (2) events emerged as a key element in the representation of data in areas such as history, cultural heritage, and multimedia. We bring together computer scientists, computational linguists and humanities and social sciences scholars in order to *build upon and expand the results in existing research communities*, e.g. NLP, Information Retrieval, Semantic Web, Social Web Analytics, Multimedia analysis, and *provide structure and deeper understanding* in history, media, journalism and cultural heritage research, with a specific focus on how events are used as a key concept for representing knowledge and organising media in online web collections. The ultimate goal is to distill a **research and application roadmaps** for events in Cultural Heritage,

e.g. achieving a social consensus on processes, identify practical standards and protocols, defining the infrastructure needed.

Our approach is two-fold following two parallel tracks. On the one hand, we dive *top-down* to *provide an comprehensive analysis of the state-of-the-art around events* and their pivotal role in enriching the content of collections in these areas. In this context, we study their added-value in enabling new meaningful interactions with multimedia collections online for humanities scholars, heritage professionals and lay audiences [4]. We also study their various aspects and potential benefits of assigning events in the representation and organisation of knowledge and media [5]. For this, we explore methods and techniques to support (1) detection, modeling and representation of events in online collections; and (2) searching, exploration and interpretation of online collections enriched with events. For example, we assess the utility of existing event models to support users in deriving useful facts.

On the other hand, we emerge a *bottom-up analysis of concrete use cases and datasets*. We guide our explorations through event detection and analysis performed by machine [6,7] and human [8,9] computation on different collections in the context of concrete use cases. We identify four groups of research questions related to (1) event identity and definitions, (2) event detection and extraction, (3) event modelling and representation, and (4) event relationships and interactions with applications.

In the context of studying the **event identity and definitions** we are interested in understanding better what is the internal structure of an event; what are the differences between events, actions and states; when two or more events the same; what are different points of view and interpretations of the same event.

To continuously improve methods and tools for **event detection and extraction** the research needs to be guided by a deeper understanding of *how events can be recognised in different media types; how can we assign a notion of novelty & veracity to events; how can we assign a level of granularity to events; how are different events related*.

To **model events and represent knowledge about events** across different domains we seek deeper understanding of shared, open or proprietary knowledge structures, such as vocabularies, taxonomies and thesauri that can build the backbone of such models. We further study how we can achieve interoperability of event structure, and what are the event representation requirements for different types of events, e.g., historical, cultural, personal events. It is also interesting to know how does Social Web influence or contribute to the understanding of event.

In this context, we move beyond the typical philosophical level discussions about events and provide the landscape of the different points of views and school of thought on that matter. To facilitate a shared and pragmatic approach to deal with events, we focus on existing models, such as the Simple Event Model³⁸, LODE³⁹, EVENT, Schema.org, Wikidata. Each of them has been developed to make use of existing vocabularies and data sources that describe events, where events refer to everything that happens, even fictional events.

Finally, we aim to understand the diversity of event relationships and their interactions with applications and data, i.e. how can events be represented in to support collection browsing, serendipitous exploration, narrative building; what are useful tools for event annotation by experts and lay crowds; what are efficient ways of crowdsourcing event annotations; what are successful methods for event visualisation & interaction.

³⁸ <http://semanticweb.cs.vu.nl/2009/11/sem/>

³⁹ <http://linkedevents.org/ontology/>

References

- [1] C Van Den Akker, A van Nuland, L van der Meij, L. Aroyo (2013). From information delivery to interpretation support: evaluating cultural heritage access on the web
- [2] Capurro, R. (2010). Digital Hermeneutics: An Outline. *AI & Society* 2010, 35 (1), 35-42
- [3] Wyatt, S., Millen, D. (Eds.) Meaning and Perspectives in the Digital Humanities. A White Paper for the establishment of a Center for Humanities and Technology (CHAT), KNAW, 2014
- [4] V De Boer, J Oomen, O Inel, L Aroyo, E Van Staveren (2015). DIVE into the event-based browsing of linked historical media
- [5] M van Erp, J Oomen, R Segers, C van den Akker, L Aroyo, G Jacobs (2011). Automatic heritage metadata enrichment with historic events. *Archives & Museum Informatics*, Toronto
- [6] Sprugnoli, R., Tonelli, S. (2016) 'One, no one and one hundred thousand events: Defining and processing events in an interdisciplinary perspective', *Natural Language Engineering*, pp. 1–22.
- [7] T Ploeger, M Kruijt, L Aroyo, F de Bakker, I Hellsten (2013). Extracting activist events from news articles using existing NLP tools and services (2013)
- [8] A Dumitrache, O Inel, B Timmermans, L Aroyo, RJ Sips (2015). CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data.
- [9] L Aroyo, C Welty (2013). Harnessing disagreement for event semantics. In the proceedings of Detection, Representation, and Exploitation of Events in the Semantic Web.

Panel: Strategies for integrating Digital Humanities skills and practices in the Humanities Curriculum

Susan Aasman (University of Groningen)⁴⁰

Stefania Scagliola (University of Luxembourg)⁴¹

In this panel we intend to evaluate policies and strategies that have been applied to integrate both DH skills and practices in Humaniora curricula. Most teaching on DH is offered in separate Minor and/or Master programs and is scarcely integrated in the regular curriculum. This reflects the skepticism with regard to the status of digital approaches to humanities research: are they supposed to gradually merge into the regular humanities curriculum or is Digital Humanities going to remain a distinctive field? (Reid, 2012). Overall, a lack of consensus on the level of expertise that should be taught can be discerned. Should they be trained to be able to make an educated choice in their future job of which kind of technological expertise they should seek? Or should the goal be to use the tools themselves and be able to customize them to their specific needs? Even teaching the very basic skills to students on how they can search, access, process, analyse and create information with digital tools, requires literally more space and time than is available within a traditional subject such as 'methods of research'. Nor does the introduction to the services of the Library that is offered yearly at the start of a humanities bachelor, suffice to cover all necessary skills (Ferrari et al, 2014, Clement, 2012). This knowledge gap is remarkable, considering the widely shared belief that DH skills are an important asset for increasing the chances of students on the job market (Clement, 2012, Scagliola et al, 2014). It is clear that the future of Digital Humanities teaching faces a number of institutional, political, logistical and pedagogical challenges.

Our intent is to offer an alternative to the usual ideal typical agendas on what should be done to solve this problem. We intend to gather best practices through the active involvement of the attendants of the workshop. We will start with a short overview of existing DH teaching courses within the Benelux, that can be retrieved through the DARIAH/CLARIAH web based [Course Registry](#). Following the short overview, three exemplary use cases from our own teaching practice will be introduced.

Case1: Integration in curriculum: Masterprogram Digital Humanities in a faculty of Arts

- Context: designing a Master program for a Faculty of Arts (History, Art History, Journalism and Media Studies, Literature, Film, European Languages and Culture, Communication Science, Information Science and Archeology), which is open to all students with a BA in one of the Arts
- Goal: offering an all-round program that combines theoretical reflection on Digital Humanities and the role of digital data in contemporary culture and society (including Art), to skill courses (coding for Humanities, creating a database) and data handling (creating/analyzing/visualizing)
- Credits: 60 EC program, no entry requirements
- Obstacle: making the shift in just one year from a regular Bachelor program based on the traditional hermeneutic framework, to more quantitative approaches requiring new skills, new methods while keeping close to the disciplinary background. This poses dilemmas on what to leave out and include.

⁴⁰ Susan Aasman is media historian and works at the Department for Media Studies at the University of Groningen. She also coordinates the Master program Digital Humanities and is Director of the Groninger Centre for Digital Humanities.

⁴¹ Stefania Scagliola is a historian and works as a postdoc at C2DH, the Centre for Contemporary and Digital History of the university of Luxembourg. She is developing a platform for teaching Digital Source Criticism.

The second case is a course that is in preparation.

Case 2: Integration in curriculum: subject Digital Source Criticism in a traditional faculty of History

- Context: Designing a platform for Digital Source Criticism for bachelor and master students to teach digital history, within a traditional faculty of history.
- Goal: teaching students the practical and theoretical implications of historical sources in digital form, teach them how to create a digital object/exhibit/publication.
- Credits: to be decided
- Obstacle: the amount of time to train skills and to create a digital object, is not available within the existing curriculum. DH is method oriented, whereas most history classes are thematic.

Case 3: Integration in the educational resources of a humanities curriculum of the Digital Humanities Course Registry.

- Context: a search environment has been designed that offers an overview of DH courses that can be taken up in the Benelux
- Goal: The goal is to offer students and lecturers the opportunity to get an overview of the courses that are taught. Students can orientate themselves and choose a bachelor, master or subject, lecturers with interest in taking up DH in their teaching can orientate themselves with regard to content and approach by drawing on the effort of their peers.
- Credits: not applicable
- Obstacle: The resource has been created, but is not integrated into the standard educational resources that are offered to students and lecturers to orientate themselves. Reaching out to the intended audience is problematic.

After a short introduction, the participants will then be divided in three groups and each group will be given a case study with an assignment related to the challenges that the case poses. They will be requested to brainstorm on possible solutions and document their suggestions in a collective online document. This will form the basis for a broader collective document that can be crowdsourced within the teaching community, turned into a publication and presented at a next Benelux DH gathering. After the brainstorm, each group presents its findings.

Our expectation is that the focus on a concrete teaching practice by scholars directly involved in the field, will yield useful insights into their best practices and strategies for expanding the interest in DH. One of the central issues remains the question whether Digital Humanities should be considered as an entity in itself competing with regular subjects or whether it should be integrated into the regular curriculum and become a standard practice.

Literature

Anusca Ferrari, Barbara Neža Brečko, Yves Punie, 'DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe', in: *eLearning Papers* 38, May 2014 – www.openeducationeurope.eu/en/elearning_papers.n.38

T. Clement (2012), 'Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind'. In Hirsch, B. (ed), *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge, U.K.: Open Book Publishers. 365-388.

R. Reid (2012), 'Graduate education and the ethics of the Digital Humanities', in: Matthew K. Gold (ed), *Debates in Digital Humanities*, Minnesota, USA.: University of Minnesota.

S. Scagliola, F. Maas, E. Stronks, 'The Teething Troubles of Teaching Digital Humanities: Sharing knowledge and mapping challenges', presentation at the DH Benelux 2014.

1. Is the Europe of Knowledge the talk of the town? Exploring the potential of digital data on MEP speeches in the European Parliament

Martina Vukasovic^{1, a}, Julie M. Birkholz^{1, b}, Jelena Brankovic^{1, 2, c}

¹ Centre for Higher Education Governance Ghent (CHEGG), Department of Sociology, Faculty of Political and Social Sciences, Ghent University, Korte Meer 5, 9000 Gent, Belgium

² Universität Bielefeld, Faculty of Sociology, Gebäude X C2-201, Bielefeld, Nordrhein-Westfalen, DE 33501

^a Corresponding author: martina.vukasovic@ugent.be, +32-9-264-8437

^b julie.birkholz@ugent.be, +32-9-264-9174

^c jelena.brancovic@ugent.be, +49-521-106-12978

Abstract

We explore whether and how increasing competences of the European Parliament (EP) across policy areas impacted its approach to higher education (HE). Using a new digital dataset containing more than 10,000 speeches delivered in the EP plenary between 2000 and 2014, we identify that total number of speeches did increase over time, particularly during the adoption of action programmes in the area of HE and related budgetary decision. HE was less referred to in the EP speeches as a stand-alone issue than in relation to other policy areas in which the EU has strong jurisdiction. We also provide tentative evidence that the variance in whether a member of the EP (MEP) speaks about HE is more linked to MEP's country of origin than party affiliation. Promises and pitfalls of digital data analysis and possible avenues for further research are also discussed.

Keywords: higher education; policy; European Parliament; Europe of Knowledge; digital data collection; semi-automatic content analysis

Acknowledgements

We would like to thank colleagues at the CHEGG (in particular Marco Seeber) and participants at CHER 2015 conference. Any remaining errors are our own. *Acknowledgements to Talk of Europe staff, editors and reviewers to be added.*

Funding

This work was supported by the Research Foundation – Flanders (FWO), under Grant G.OC42.13N.

Introduction

Since 2000, the European Union (EU) has put knowledge at the centre of its strategic endeavours. The aim of the Lisbon Strategy was for Europe to become the most advanced knowledge-based economy in the world by 2010. As a result, during the 2000s, the European Commission put forward several communications focusing on the role of universities in this process and the necessity for a university modernization strategy (e.g. European Commission, 2006), culminating with the Europe 2020 in which knowledge is essential for ensuring smart, inclusive and sustainable growth (European Commission, 2010). Throughout this period, knowledge has been 'exported' to other policy areas as a policy solution (Elken, Gornitzka, Maassen, & Vukasović, 2011), while the funding of EU programmes fostering cooperation in this area has been increasing, despite the financial crisis – e.g. for the 2014-2020 period, there is a 30% increase of funds allocated to research cooperation, and 40% increase for education.

Although these developments have been the focus of many studies,⁴² most of them are concerned with the creation of specific institutions (e.g. the European Institute of Technology) or the Bologna Process and its relationship with the EU initiatives, whereby they typically highlight the individual policy entrepreneurs or the role of the European Commission (and sometimes also the European Court of Justice, ECJ). Other EU institutions, in particular the European Parliament (EP) and its involvement in policy coordination in this area, have received far less attention thus far, which reflects neither the importance of HE for the whole European project nor the increasing importance of the Parliament in EU decision-making (see below).

With this in mind, the present study focuses on the extent and the manner in which HE has been discussed in the EP since 2000, exploring a new digital dataset containing MEP speeches delivered during the period studied. We start by outlining the changes in how EU approaches the topic of HE and the overall role of the EP in EU decision-making. From this, we derive a set of expectations concerning how HE is discussed in the EP which we investigate through an exploratory research design. Namely, we undertake digital data collection methods and semi-automatic content analysis coding on a set of more than 10,000 speeches given in the EP between January 2000 and December 2014, identified through a set of search terms using 'The Talk of Europe' dataset.⁴³ We analyse the data and identify patterns related to 'when, who and how' speaks about HE in the EP. We then discuss our findings, as well as the promises and pitfalls of digital data analysis as a method, and offer some directions for future research.

Context and analytical patterns of interest

HE and the EU

In the EU context, HE has largely been considered a specialised policy area steered by national ministerial administrations and strongly influenced by expert committees and local sectoral interests. Education in general has long remained an area of national competence (Gornitzka, 2009), meaning that the legislative bodies of the EU (the EP and the Council) do not have regulative competences in the area of HE. Previous to the Treaty of Lisbon, this was reinforced in the principle of subsidiarity – decisions were taken at the lowest possible governance level, in this case the national authorities. From 1 December 2009 onwards, when the Treaty of Lisbon came into force, education has been considered as a supporting EU competence allowing 'the Union to carry out actions to support, coordinate or supplement Member States' actions' in this area. This change potentially provides more leeway to the EU in this domain, although EU still cannot engage in these actions on its own. However, a number of caveats need to be addressed.

First, interest in European level policy coordination in the area of HE has existed since the early days of the European project. As Corbett (2005) states, ever since the European Coal and Steel Community, European level policy entrepreneurs in various EU institutions have been pushing for different European initiatives targeting education. Their efforts eventually laid the ground for the Erasmus programme and a number of pilot projects focusing on policy coordination, such as cooperation in the area of quality assurance (ENQA 2010; EU Council 1998). This trend has been strengthened by several rulings of the ECJ concerning recognition of qualifications (see e.g. Corbett 2005 on Gravier decision), as well as regulation concerning recognition of qualifications, in particular for regulated professions (Beerens, 2008). There are also indications that HE (and research) may increasingly become subject to EU primary law (i.e. EU level regulation) concerning competition (an

⁴² See e.g. Amaral et al. (2009), Chou and Gornitzka (2014), Corbett (2005), Huisman and de Jong (2014), Maassen and Olsen (2007).

⁴³ <http://www.talkofeurope.eu/data/> (accessed 22 January 2017).

exclusive competence of the EU after the Treaty of Lisbon), due to the blurring of the distinction between public and private aspects of HE (Gideon, 2015).

Second, (higher) education, research and innovation comprise the so-called knowledge triangle which has been the cornerstone of the EU's strategic documents since the Lisbon Summit in 2000. The focus on developing the EU as a Europe of Knowledge or Innovation Europe has remained strong. Thus, one could argue that integration in this area can be considered a *sine qua non* of European integration as such (European Commission, 2010). HE is being 'exported' to other policy areas – economic competitiveness, social cohesion, environment, security, foreign relations etc. – as a policy solution and modernization of HE is seen as a key ingredient of political, social, economic and cultural development (Elken et al., 2011). Due to this functional 'spill-over' from areas in which the EU does have formal regulative competences, this means that HE is becoming a topic of growing interest for EU institutions.

Third, in the area of HE the EU has been employing the so-called Open Method of Coordination (hereinafter: OMC). OMC relies on voluntary setting of standards and benchmarks, and includes development of procedures designed to monitor progress. While this approach may, at first glance, seem rather soft given its voluntary nature and ample room for window-dressing, evidence suggests that the possibility inherent in the OMC to 'name and shame' laggards can actually be a powerful instrument leading to significant changes on both the national and institutional level (Gornitzka, 2014). The fact that these changes do not necessarily result in clear and deep convergence is less an indication of OMC's softness and more an indication of the complexity of implementation processes in HE (Musselin, 2005).

In sum, the EU has been increasingly focusing on policy coordination in the area of HE, either on its own or due to spill-over from other policy areas in which it has explicit competences. While most of the activities in this area have been led by the executive branch – the European Commission (EC), other EU institutions have focused on HE as well, including the EP which, amongst other, is tasked with oversight of the EC.

The role of the European Parliament in the EU decision-making

Overall, in the case of EU decision-making, the distribution of power is assessed as rather complex (Börzel, 2010). While the executive, judicial and legislative powers in the EU are shared, the basic distinction between government, courts and the parliament that exists on national levels does not exist in the same way on the European level. This is in particular the case for legislative competences which are currently shared between the EP and the Council, a set-up referred to as 'co-decision'.

The specification and division of tasks between the different EU institutions has been evolving since the very beginning of the European integration project and this is in particular true for the "legislative powers of the EP [which] have grown sequentially" (Pollack, 2010, p. 31). The seeds of EP's empowerment can be found already in the treaties of 1970 and 1975 which gave the EP some control over the EU budget and introduced also the Court of Auditors. The Single European Act from 1986 also gave the EP increased legislative power and expanded the overall EU policy scope, extending and deepening EU's competences in more areas (Wallace, Pollack, & Young, 2010). Co-decision between EP and the Council – implying that a decision needs to be accepted by both bodies – was first introduced in the 1992 Treaty on EU (Maastricht), while the 1997 Treaty of Amsterdam introduced strong requirements for EP's assent on enlargement and appointments of the Commission. Overall, EP's involvement in EU decision-making has evolved from a non-binding consultation procedure to a co-decision procedure with the Council in the 1990s, only to be further strengthened in the 2000s by establishing co-decision as the standard operating procedure used for majority of policy areas (Pollack, 2010).

Furthermore, the EP is tasked with approving the EU budget and discharging the accounts of the previous year (Laffan & Lindner, 2010). Concerning budget approval, these decisions are important because they are highly visible to MEPs constituents and are possibly contentious given that potential winners and losers can be clearly identified – “since it has been granted budgetary powers in 1975, the EP has regarded the EU finances as one of its key channels of influence vis-à-vis the Council” (Laffan & Lindner, 2010, p. 214). This is where the EP tries to influence decisions at both the macro level – concerning multi-annual funding frameworks, as well as the micro level – concerning specific programmes and projects. An example of the former concerns the strong focus on research and technology in the discussion of the Financial Perspective for 2007-2013 (reflecting the findings of the so-called Sapir report), where the EP also supported the EC’s proposal to strengthen expenditure for public goods, effectively positioning itself against the some Member States (Laffan & Lindner, 2010). Example of the latter is the decision concerning Erasmus Mundus budget in 2003 (Corbett, 2005) and the EP’s concern over the Juncker Commission to use 2/3 of the Horizon 2020 funding for the EU’s investment fund.⁴⁴ Given that the multi-annual budget plan actually has the status of a law, binding for several years, the EP’s deliberations and decisions on the budget issues have become even more important.

The EP also plays an important role in appointing the Commission; it approves the Commission President and has the power to hold the Commission accountable. For example, it was effectively the EP which forced the Santer Commission to resign in the late 1990s, following the claims of insufficient transparency in spending of EU funds. The EP also delayed the endorsement of the 2004 Commission due to the proposed composition and the endorsement of the 2009 Commission President. For the Juncker Commission, the MEPs put forth a number of requests concerning the portfolios of the different Commissioners and the proportion of female Commissioners before approving the overall composition.

Overall, the EP is currently in the position to constrain the agenda-setting activities of the EC and it can also explicitly ask the EC to deal with specific issues (Young, 2010). Given its role in the co-decision procedure, it can effectively act as a veto player and block decision-making (Finke, 2010). However, it has almost no involvement in the process of implementation and policy evaluation (Young, 2010). In general, since the early days of the European integration project, the EP has increased its influence on European level decision-making, where in exchange the EC and the Council have arguably lost some influence. However, a detailed analysis of inter-institutional power relationships would need to take into account that none of these institutions are unitary actors and that their own internal dynamic is important as well.

What goes on inside the EP and why is it important?

The EP is composed of MEPs, the vast majority of which are organized into European party families, while the rest are ‘non-attached’ MEPs. The candidates for MEPs run at nationally organized elections, where the number of MEPs to be elected from each state depends on the country’s population. However, once elected, the MEPs are grouped in the EP not according to their countries, but in accordance to their partisan affiliation. The composition of the EP and the total number of MEPs per parliamentary term of interest for this study is presented in Table 1.

Given that it is a supranational legislature, EP’s connection to the electorate is by some considered “notably weak” (Young, 2010, p. 58), although it should be acknowledged that the EP is effectively the only EU institution whose members are directly elected. Evidence suggests that, once in the EP, the decisions of the MEPs are more determined by the generic left-right political cleavages between

44 See e.g. <https://euobserver.com/economic/128867> (page accessed 1 March 2017).

the different European party families and their positions concerning the scope and level of appropriate European integration than by the MEPs' country affiliations (Finke, 2010; Pollack, 2010).

Table 1 – Number of MEPs across party families and parliamentary terms. Source: EP website.

Number of MEPs (per party family)	5 th term 1999- 2004	6 th term 2004- 2009	7 th term 2009- 2014 ⁴⁵
European United Left/Nordic Green Left (GUE-NGL)	42	41	35
Progressive Alliance of Socialist and Democrats (S&D); formerly PES	180	200	184
Greens/European Free Alliance (Greens/EFA)	48	42	55
Alliance of Liberals and Democrats (ALDE); previously ELDR	50	88	84
European People's Party (EPP), formerly EPP-ED	233	268	265
Europe of Freedom and Direct Democracy (EFDD), formerly IND/DEM or EFD	16	37	32
European Conservatives and Reformists (ECR), formerly UEN	30	27	54
Non-attached (NA)	9	29	27
Total	626	732	736

While preparatory work is carried out in specialist committees of the EP (Wallace, 2010), plenary sessions taking place every month in Strasbourg serve as an opportunity for MEPs to address each other, as well as other EU institutions and the public (Proksch & Slapin, 2010; Slapin & Proksch, 2010). The speeches during these sessions serve several purposes: (a) arguing in favour or against a legislative proposal, (b) scrutinizing other actors, in particular those over which the EP has oversight (e.g. the EC), (c) sending signals to national constituents, (d) other members of the party group or (e) other members of the EP (Slapin & Proksch, 2010). The sessions are sometimes structured around an opening statement or a proposal by the EC, followed by a rapporteur of the relevant EP committee (Proksch & Slapin, 2010). The latter play a particularly important role: they are the ones steering negotiations within the committees and working on ensuring the support across different political groups (Kohler, 2014). While this 'behind-the-scene' work potentially limits the possibilities for debate and conflict in the plenary session between the different party families (Kohler, 2014), rapporteur's speeches in the plenary are nevertheless important as indicators of the outcomes of negotiations within the committees.⁴⁶ After the rapporteur, the speaking time is allocated to party families, with MEPs of the largest family speaking first. Allocation of time between the MEPs within one family is done internally, and the individual speech cannot last more than three minutes. At the end of the debate and before the vote, the EC representatives may reply and indicate the EC's position on the proposal (Proksch & Slapin, 2010). Importantly, as the EP also has the power to put forward issues on its own, and not only to follow EC's agenda, MEPs can speak on a wide range of topics, both those in which the EP has explicit competences concerning regulation adoption (the so-called 'hard law'), as well as those subject to softer policy coordination. Effectively, MEPs use plenary sessions as the opportunity to give speeches both to communicate their own positions towards the

⁴⁵ The number of MEPs in the Seventh Parliamentary term changed twice, first due to the Lisbon Treaty entering into force in December 2009 (to 754 MEPs) and then due to Croatia joining in July 2013 (to 766 MEPs).

⁴⁶ The 'Talk of Europe' dataset includes only speeches made in the plenary session.

general public and their own constituents, as well as to coordinate with other actors, relying on discursive practices as instruments of change (Schmidt, 2010).

Expectations

Given that (a) HE has become increasingly important for the overall EU strategic development, that (b) HE has been exported to other policy areas as a policy solution, that (c) despite the lack of strong regulative competences, there is significant HE policy coordination at the EU-level, that (d) there has been a gradual empowerment of the EP with regards to the EU level decision-making, in particular when it comes to budget decisions, and that (e) the behaviour of MEPs in general seems to be determined more by their party affiliation than by their country affiliation, the following patterns with regards to how HE is considered in the EP can be expected:

1. The total number of MEP speeches referring to HE increases over time. The most significant increase is expected in relation to the adoption of EU action programmes and related budgetary decisions.
2. HE is more often referred to in the EP speeches in relation to other policy areas in which the EU has regulative competences, than as a stand-alone issue.
3. Whether or not an MEP makes a speech addressing HE is more strongly linked to his/her party family affiliation than to the country of origin.

Data and method

To investigate the role of HE in the EP we studied speeches delivered in the EP plenary using the ‘Talk of Europe’ – a linked open data infrastructure (van Aggelen, Hollink, Kemman, Kleppe, & Beunders, 2016), which comprises speeches given in the EP from 1999 – 2014 (translated into English) and related data available through the European Data Portal.⁴⁷ ‘Talk of Europe’ allows the use of semantic queries to retrieve data stored in the Resource Description Framework (RDF), a computer data language (Juric, Hollink, & Houben, 2012). By formally linking traditionally distributed datasets, queries can be implemented to identify specific artefacts and related meta-data. These digital provisions offer a number of advantages to the researchers wishing to investigate this data. First, one can automatically identify a large amount of documents in a straightforward manner, as opposed to querying one database of MEPs’ speeches, querying another to retrieve data about dates, agenda items, MEPs, etc. and then merging them. In addition, instead of manually inferring possible connections, the data is automatically linked and compiled as one artefact, which saves substantial time (e.g. that John Smith in one database is the same in the other). Depending on the query size, querying such data may be run in minutes, if not seconds.

However, with the advent of these tools come challenges associated with designing, conducting, and interpreting research results (Bar-Ilan, 2001). As the data is highly sensitive to the query commands, the specifics of the query influence the data returned. Thus the researcher has to be acquainted with the database, its possibilities and the nature of the request, so as not to jeopardize validity of the design. For example, the researcher must know whether the database is continuously being updated, what characteristics are available of the artefacts being queried, and how the data needs to be structured to conduct the appropriate analysis given the specific research design. Taking into consideration the nature of such data and the focus on the development of a query to identify specific, crafted valid samples within these relatively large datasets, a total description of the entire dataset is rarely feasible nor conducted. This implies that normalization – comparing an entire dataset to the selected sample – for the purposes of verifying the representativeness of sample (a standard in quantitative research) is in this study conducted using other means (see ‘Results and Discussion’). Although it may question some assumptions of specific designs, for example procedures

⁴⁷ <https://www.europeandataportal.eu/> (page accessed 1 March 2017).

and characteristics necessary for inferential statistics, we would argue that such issues do not warrant disregarding such data, but rather require that these specificities are transparently presented and openly discussed, as we work towards building a methodological toolkit fit for analysing such digital data. Despite these pitfalls, we contend that exploring such data would offer valuable and perhaps unique insights about phenomena studied.

In analysing these data we take an exploratory approach. We started our research by developing a set of terms related to HE, which were then reviewed by a number of HE researchers (see Appendix for the list of all terms queried). In order to identify speeches, a query, using these key words, was developed by the second author, with the assistance of the ‘Talk of Europe’ team. This returned the speeches in a text format, as well as the related meta-data (if available) of the: *title of the speech*, *date of the speech*, *URL to the original speech*, *identification of the speaker*, *speaker’s country affiliation*, and the *speaker’s party affiliation* (if known or applicable). Querying these words resulted in a set of 10,180 unique speeches (all including at least one HE term from our list, duplicates removed) and related meta-data representing all potential discussions on HE in the EP since 1999. The meta-data – which constitute essentially textual data – were coded in order to allow for a systematic analysis of temporal, topical and country/party affiliation patterns. Importantly, the query developed focused on identifying specific terms used in speeches, not a description of the entire dataset of speeches. This approach mimics techniques implemented in other studies using this dataset (van Aggelen et al., 2016). In addition to this data, we used publicly available information on the number of MEPs per country or per party family during the period studied. The treatment of the data is presented in the Table 2.

Table 2 – Variables and treatment. Source: Authors.

Variable	Type of data	Treatment
Title of the speech	Textual data	Manually coded data in relation to the topics, see Table 3.
Date of the speech	Date	n/a (not treated here)
Unique ID of the speaker, given by EP	Nominal	n/a (not treated here)
Speaker’s country of affiliation	Textual data	Coded to nominal data
Speaker’s party affiliation, if known	Textual data	Coded to nominal data

To efficiently identify topics according to key words in the speech titles, as presented in Table 3, speeches were semi-automatically coded-using both manual and computational coding (Lewis, Zamith, & Hermida, 2013). This resulted in one code per speech, following a hierarchical schema: with a title containing one of the HE terms identified earlier taking primacy, then a non-HE topic (e.g. geographical determinant, demographic determinant or references to other policy sectors). For example, if the title refers to Roma or the Danube Region but the speech mentions a HE term, it constitutes a non-HE topic where HE has been discussed in related to another policy topic. Speeches that included ‘vote’, ‘budget’ or reference to a procedural matter were also coded into separate categories (see Table 3 for details). Voting and budget formally represent two different activities, whereby one is possibly a specific discussion of the budget, compared to a discussion on the voting itself as a decision-making process of the EP. We acknowledge that in a small number of cases these may overlap, given the EP often votes on budgets. All other formal activities related specifically to procedure and the organization of the EU in general are considered as procedural topics.

In order to explore these patterns, these coded data, together with the above mentioned meta-data of text origin (string variables), were transformed into nominal categorical variables. The dataset was then used to explore a) the temporal patterns of the use of HE terms in speeches over time, b) the

topical patterns of the use of HE terms in the different types of speeches, c) the role of the country and party in explaining the use of these terms over time and in specific topics.

Table 3 - Coding scheme. Source: Authors.

Code	Description
HE	Speeches with a title that included one of our key words and addressed HE as the specific topic
Non- HE	Speeches with the mention of a geographical place in the title (e.g. country, city or region), or a specific group of people in the title (e.g. women, youth, disabled, elderly, Roma), or an issue that is not specifically related to HE (e.g. economy, human rights, employment, labour, resources, security, environment, defence, transportation, and so forth).
Vote	Speeches with the title Vote or Votes
Budget	Speeches with the mention of the word budget in the title
Procedural	Speeches with a mention in the title on procedural matters of the EU itself (e.g. review of EC notes, announcements, and so forth)
Unidentified	Speeches that are not attributable to a topic given the lack of detail in the title

We acknowledge a number of limitations to our design. The reliability of the public data is related to the accuracy of EU Open Data Portal and the ‘Talk of Europe’ infrastructure in both publishing and accurately linking related data. Given the bigger nature of this data, it is expected that less significant ‘bugs’ may occur, but that this ‘noise’ would be systemic and thus would not significantly influence results. In this respect, we have encountered an unprecedented amount of unattributed party affiliations in the 7th EP session which reflected missing data. Thus, to ensure validity, in considering the extent to which MEP’s speaking on issues related to HE is determined by his/her country of origin or party family affiliation we have not analysed the 7th term. Within the available data we were not able to confirm whether all speakers were MEPs, or guest speakers, although we could safely assume that the number of non-MEPs speaking in the EP plenary sessions is very low and thus not significant in a way that could distort our findings.

Results and discussion

As previously indicated, our query retrieved a total of 10,180 speeches containing one or more of the terms in our ‘dictionary’ (see Appendix). Given that the output of the query does not contain a list of terms that were found in a particular speech, it was not possible to systematically measure the co-occurrences of the terms across all of the 10,180 speeches and to use such data to test the sensitivity of the query to the content of the ‘dictionary’.

Given these limitations, we have devised an alternative approach. We focused on the potentially most problematic terms in the ‘dictionary’, i.e. terms that may appear in speeches with no linkage to HE whatsoever: innovation, mobility, science, technology, and training. We have queried the ‘Talk of Europe’ infrastructure for these five terms separately and analysed the overlap between speeches retrieved this way and speeches retrieved when querying for two terms definitely linked to HE – ‘higher education’ and ‘university’. The results are presented in Table 4.

Table 4 – Number of speeches containing one or more of the selected terms. Source: Authors.

	X				
Number of speeches...	innovation	mobility	science	technology	training
A: containing one of the terms (X)	923	752	534	973	1011
B: containing X AND 'higher education' (Y)	262	269	180	264	281
C: containing X AND 'university' (Z)	370	322	258	421	403
containing (X AND Y) OR (X AND Z) = B+C	632	591	438	685	684
D: containing Y AND Z	221	221	221	221	221
containing (X AND NOT Y) OR (X AND NOT Z) = B+C-D	512	382	317	509	548

Thus, there are 2,268 speeches (22.2% of the total number of speeches in our dataset) that contain at least one of the five problematic terms (X), but do not contain 'higher education' (Y) or 'university' (Z), i.e. potentially there are 22.2 % of speeches in the dataset that should not be there. However, we need to stress that this is actually the maximum possible value, for two reasons: (1) we just explored the co-occurrence of the problematic terms (X) with two other terms ('higher education' and 'university') and not with other terms in the 'dictionary' which may also be closely linked to HE (e.g. student, academic); and (2) we ignored the possibility that there may be co-occurrences of the different Xs in the same speech (e.g. 'innovation' and 'technology') and merely added the different numbers in the last row of Table 4. Notwithstanding that the actual proportion of speeches that do not belong in the dataset is very likely lower than 22.2%, we will proceed with our exploration of the data taking this into account.

In relation to our expectation that the total number of MEP speeches mentioning HE increases over time, Figure 1 presents the frequency of such speeches for the 5th, 6th and 7th term (aggregated for a four-month period).

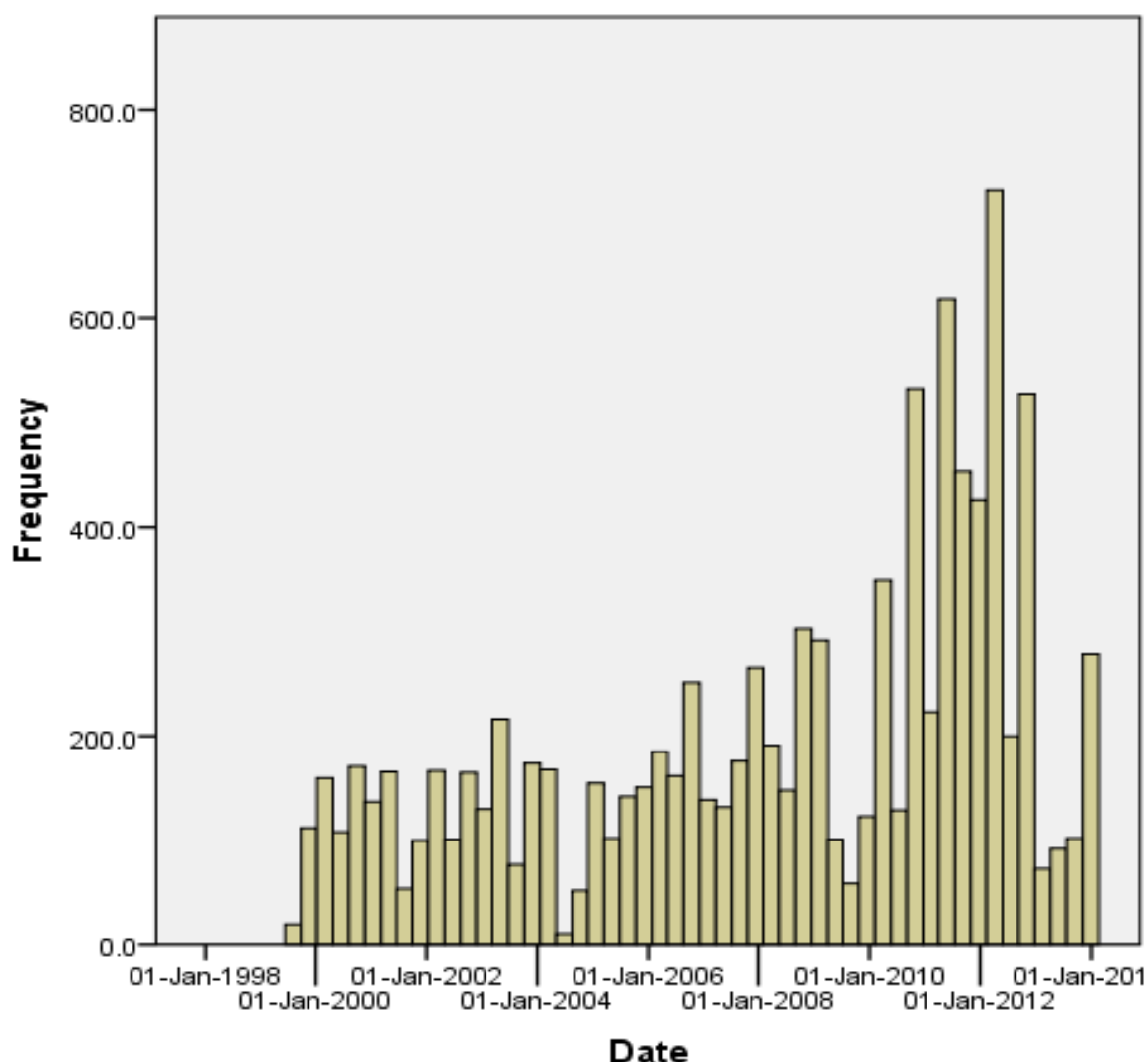


Fig. 1 - Speeches referring to HE, over time. Source: Authors.

As Figure 1 shows, there is an increase in the number of speeches containing at least one of the terms we identified as being attributed to HE over time. The figure also helps us identify moments of increased frequency, such as the end of 2008, parts of 2011 or the end of 2013.

A closer look into the dataset reveals that these increases are related to the activity around the adoption of specific programmes and decisions concerning HE, such as:

- the 'Erasmus Mundus II' programme – 31 speeches on this topic on 20 October 2008 alone;
- European Quality Assurance Reference Framework for VET – 38 in December 2008;
- the report on the 'Youth on the Move' (which also includes student mobility programmes, such as, currently, Erasmus +) – 57 speeches in May 2011;
- Agenda for new skills and jobs – 53 speeches in October 2011;
- Modernising Europe's Higher Education Systems – 38 in April 2012;
- a debate titled 'Is Erasmus in danger?' – 42 speeches in October 2012;
- 'Erasmus +' programme under 'Erasmus for All' item on the agenda – 142 speeches in November 2013.

This also means that the increased frequency cannot be due to the 22.2% potentially problematic speeches in our dataset. At the same time, moments of lowest frequency pertain to the transition between different parliamentary terms.

Concerning the context in which HE is referred to, in the majority of speeches HE is not discussed as a stand-alone issue but rather in relation to other policy issues or in relation to the vote explanations (Table 5). These other policy issues include areas that could be considered closely related to HE, such as general education, youth issues or recognition of professional qualifications, but also include areas that can be considered as rather distant from HE, e.g. visa issues, arms sales, maritime policy etc. The high proportion of the speeches categorized under the ‘Vote’ topic indicates that HE related terms are also referred to during explanations of voting procedures as well as discussions concerning implications of the votes.

Table 5 – Distribution of speeches referring to HE in relation to their main topic. Source: Authors.

Main topic	Number of speeches	% in relation to all speeches including HE terms
Non-HE	3,844	37.76%
Vote	3,131	30.76%
HE	1,358	13.34%
Procedure	1,122	11.02%
Budget	713	7.00%
NI	12	0.12%

While the structure of our dataset does not allow for a more refined analysis with regards to how HE is referred to in relation to other policy issues or voting, it is clear that HE does not feature prominently as a stand-alone issue but that it is most often referred to in relation to other policy issues in which the EU has regulatory competences, even when taking into account that potentially 22.2% of the speeches – all of which would be on topics other than HE – perhaps should not be in our dataset.

Concerning our third expectation, we focused only on the 5th and 6th term for which we had clear party family affiliation for each MEP and restructured the dataset so the MEP (and not an individual speech) is the data unit. We then calculated for each MEP the proportion of speeches that had HE as its main topic in relation to the total number of speeches given by said MEP (hereinafter: HE speeches), and based on this explored the variance in proportion of HE speeches in SPSS with a two-way ANOVA using country and party as fixed factors (Field, 2009). The results show that a statistically significant difference in the proportion of MEP speeches that are on HE exists only for country affiliation (and that only at $p < 0.05$ level of significance) while the difference for party affiliation is not significant. This can be considered as a suggestion that the country of origin is more strongly linked to the variance in proportion of speeches an MEP makes that have as their main topic HE, though primarily a tentative one, given the potential that a certain number of speeches – likely less than 22.2% because this analysis concerns only 5th and 6th term – should not be considered in this dataset.

Conclusions

The findings presented in this paper are the result of the first exploration of the ‘Talk of Europe’ dataset. They suggest that Europe of Knowledge is becoming the talk of the town in the European Parliament. The total number of MEP speeches, either specifically dedicated to HE or mentioning HE in speeches dedicated to other issues, appears to have increased over time, particularly during the adoption of EU action programmes in the area of HE and related budgetary decision. Moreover, over the period analysed, HE was less referred to in the EP speeches as a stand-alone issue than in relation

to other policy areas in which the EU has strong jurisdiction. Finally, the tentative findings indicate that the variance in whether an MEP speaks about HE is more linked to country of origin than party affiliation.

More generally, these findings attest to the increasing role of the EP in HE policy making, which has been largely overlooked in studies on European level dynamic in HE. As this study demonstrates, a closer look at the EP potentially offers a wealth of information on how HE, both in relation to other areas and as a policy issue in its own right, is considered and decided upon by the only directly elected institution of the European Union. Available databases, such as the 'Talk of Europe', but also the rich repositories of publicly available information on the European Union websites, therefore, offer a promise of a better insight into these matters.

In this study we have also tried to explore the potentials and limitations of using big(ger) digital data. We believe we have accurately shown how such methods and analysis can be useful in policy research, while also trying to highlight some of their shortcomings. Regarding the latter, the choice of research design did not allow for traditionally accepted methods of analysis to be employed, together with the assumptions necessary to conduct those analyses. An example of this would be the difficulties with developing a valid query from a frequently changing dataset, which in turn prevents normalization of data. We argue, and have shown in this research, that such difficulties should not automatically mean that such data is of no use but rather that the methods distinct to digital data need to be employed and transparency needs to be ensured. Specifically, to this project, we acknowledge a limitation of the developed query to individually identify terms where co-occurrence can be assessed. In developing future queries using the 'Talk of Europe' infrastructure one should attempt to build a data set that would allow a query to consider an additional number of characteristics, at least: (1) the total number of speeches at the date of data of collection (to enable normalization), and (2) data on co-occurrence of different terms of interest. These two additions would allow for further quantitative analysis, but also more confidence in claiming that certain mechanism(s) and relationships are at play, which we can only now provide as tentative findings.

Taking into account the abovementioned measures concerning the data infrastructure, a number of possible avenues for further research become open. First, in-depth analysis of the textual data contained in selected speeches would allow for further exploring the content of these speeches, e.g. what preferences and positions are MEPs putting forward and how this may change over time. Moreover, relationships between the EP and other EU institutions, such as the European Commission and the Council of the EU can be analysed by, for example, analysing the extent to which MEPs refer to HE when responding to initiatives of other EU institutions compared to speaking about HE without an external prompt. The initial explorations presented in this paper can thus serve as the backdrop for further in-depth analysis of MEP behaviour concerning higher education using digital data.

References

- Bar-Ilan, J. (2001). Data collection methods on the Web for infometric purposes — A review and analysis. *Scientometrics*, 50(1), 7-32. doi:10.1023/a:1005682102768
- Beerkens, E. (2008). The Emergence and Institutionalisation of the European Higher Education and Research Area. *European Journal of Education*, 43(4), 407-425. doi:10.1111/j.1465-3435.2008.00371.x
- Börzel, T. A. (2010). European Governance: Negotiation and Competition in the Shadow of Hierarchy. *JCMS: Journal of Common Market Studies*, 48(2), 191-219. doi:10.1111/j.1468-5965.2009.02049.x
- Chou, M.-H., & Gornitzka, Å. (Eds.). (2014). Building the knowledge economy in Europe: New constellations in European research and higher education governance. Cheltenham: Edward Elgar.

- Corbett, A. (2005). *Universities and the Europe of knowledge: Ideas, institutions and policy entrepreneurship in European Union Higher Education Policy, 1955–2005*. Basingstoke: Palgrave MacMillan.
- Elken, M., Gornitzka, Å., Maassen, P., & Vukasović, M. (2011). *European integration and the transformation of higher education*. Oslo: University of Oslo.
- European Commission. (2006). *Delivering on the modernization agenda for universities: Education, research and innovation*. Brussels.
- European Commission. (2010). *Europe 2020: A strategy for smart, sustainable and inclusive growth*. (COM(2010) 2020 final). Brussels: EC.
- Field, A. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)*. Los Angeles: SAGE.
- Finke, D. (2010). *European integration and its limits: intergovernmental conflicts and their domestic origins*. Colchester: ECPR Press.
- Gideon, A. (2015). The Position of Higher Education Institutions in a Changing European Context: An EU Law Perspective. *JCMS: Journal of Common Market Studies*, n/a-n/a. doi:10.1111/jcms.12235
- Gornitzka, Å. (2009). Networking Administration in Areas of National Sensitivity: The Commission and European Higher Education. In A. Amaral, G. Neave, C. Musselin, & P. Maassen (Eds.), *European Integration and the Governance of Higher Education and Research* (Vol. 26, pp. 109-131): Springer Netherlands.
- Gornitzka, Å. (2014). How strong are the European Union's soft modes of governance? The use of the Open Method of Coordination in national policy-making in the knowledge policy domain. In M.-H. Chou & Å. Gornitzka (Eds.), *Building the Knowledge Economy in Europe: New constellations in European Research and Higher Education Governance* (pp. 160-187). Cheltenham: Edward Elgar.
- Huisman, J., & de Jong, D. (2014). The Construction of the European Institute of Innovation and Technology: The Realisation of an Ambiguous Policy Idea. *Journal of European Integration*, 36(4), 357-374. doi:10.1080/07036337.2013.845179
- Juric, D., Hollink, L., & Houben, G. J. (2012). *Bringing parliamentary debates to the Semantic Web*. Paper presented at the 11th International Semantic Web Conference, workshop on Detection, Representation and Exploitation of Events in the Semantic Web (DeRIVE 2012), Boston.
- Kohler, M. (2014). European Governance and the European Parliament: From Talking Shop to Legislative Powerhouse. *JCMS: Journal of Common Market Studies*, 52(3), 600-615. doi:10.1111/jcms.12095
- Laffan, B., & Lindner, J. (2010). The Budget. Who Gets What, When, and How? In H. Wallace, M. A. Pollack, & A. R. Young (Eds.), *Policy-Making in the European Union* (pp. 208-228). Oxford: Oxford University Press.
- Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, 57(1), 34-52. doi:10.1080/08838151.2012.761702
- Maassen, P., & Olsen, J. P. (Eds.). (2007). *University Dynamics and European integration*. Dordrecht: Springer.
- Musselin, C. (2005). Change or Continuity in Higher Education Governance? In I. Bleiklie & M. Henkel (Eds.), *Governing knowledge: A study of continuity and change in higher education* (Vol. 9, pp. 65-79). Dordrecht: Springer Netherlands.

- Pollack, M. A. (2010). Theorizing EU policy-making. In H. Wallace, M. A. Pollack, & A. R. Young (Eds.), *Policy-Making in the European Union* (pp. 15-44). Oxford: Oxford University Press.
- Proksch, S.-O., & Slapin, J. B. (2010). Position Taking in European Parliament Speeches. *British Journal of Political Science*, 40(03), 587-611. doi:doi:10.1017/S0007123409990299
- Schmidt, V. A. (2010). Taking ideas and discourse seriously: explaining change through discursive institutionalism as the fourth 'new institutionalism'. *European Political Science Review*, 2(01), 1-25. doi:doi:10.1017/S175577390999021X
- Slapin, J. B., & Proksch, S.-O. (2010). Look who's talking: Parliamentary debate in the European Union. *European Union Politics*, 11(3), 333-357. doi:10.1177/1465116510369266
- van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2016). The debates of the European Parliament as Linked Open Data. *Semantic Web*(Preprint), 1-10.
- Wallace, H. (2010). An Institutional Anatomy and Five Policy Modes. In H. Wallace, M. A. Pollack, & A. R. Young (Eds.), *Policy-Making in the European Union* (pp. 69-104). Oxford: Oxford University Press.
- Wallace, H., Pollack, M. A., & Young, A. R. (2010). An Overview. In H. Wallace, M. A. Pollack, & A. R. Young (Eds.), *Policy-Making in the European Union* (pp. 3-13). Oxford: Oxford University Press.
- Young, A. R. (2010). The European Policy Process in Comparative Perspective. In H. Wallace, M. A. Pollack, & A. R. Young (Eds.), *Policy-Making in the European Union* (pp. 45-68). Oxford: Oxford University Press.

Appendix – HE terms queried

(Note: the 'Talk of Europe' database can query terms consisting of one or two words)

Academia	polytechnic
Academic, academics	Quality assurance
bachelor, bachelors	science
Bologna Process	Skill and skills
Copenhagen process	Socrates
COST	STEM
Curriculum	Student and students
diploma supplement, diploma supplements	technology
ECTS	Tempus
EHEA	tertiary education
employability	Training
Erasmus	university
Erasmus Mundus	VET
Erasmus+	vocational training
European Institute (to identify European Institute of Technology)	
European Standards (to identify European Standards and Guidelines)	
European University	
Framework Programme	
Graduate, graduates	
higher education	
Horizon 2020	
innovation	
knowledge (to identify knowledge-based economy)	
learning (to identify lifelong learning issues and Lifelong Learning Programme)	
LLP	
master, master's	
Mobility	

2. Mining meaning. From network analysis to algorithmic semantic data mining Introduction

Paul Verhaar and Mirko Tobias Schäfer

Bringing digital methods to linguistics, our paper uses a data set consisting of tweets about the refugee crisis as a baseline for semantic analysis. This paper describes how the analysis of a corpus of status messages can lead to the definition of linguistic fingerprints for detecting ideological positions. We use a network analysis of retweets to reveal the different positions of participants within the highly polarized debate. Mapping the network returns two opposing clusters in the debate; one expresses a mildly positive stance towards refugees, considering refugees to be victims, and is up to a certain extent welcoming them. The other one is opposed to refugees, portraying them as criminals or profiteers. The clusters can also roughly be divided in political preference; left wing versus right wing. As the political differences are obvious, we use this as a baseline for further analysis based on the content of tweets. Mining the timelines provides insights into the distinctive use of language by the participants of the opposed clusters. Our paper describes a general method for analysing a corpus of status updates in order to identify ‘linguistic fingerprints’ revealing ideological positions.

Linguistics meets digital methods

Our approach combines digital methods for Twitter analysis with linguistic methods. This means that we analyse the structure of the debate and consider how language plays a role in carrying keys for identifying ideological positions within the debate. Analysing Twitter or social media messages is not new. Previous research has focussed on email analysis (Groh and Hauffa, 2011), network analysis on retweet behaviour (Passman et al., 2014), abstracting personality from social media (Schwartz et al., 2013) and style accommodation (Danescu-Niculescu-Mizil et al., 2011). We are not aware of a study combining digital methods for network analysis with linguistic analysis to breed a connection with a solid baseline. Many concepts in new media studies and linguistics can be combined. Register is an important part of our everyday conversation, useful for research on the online domain. People tend to convey their messages in different ways and forms depending on the person they share information with (Danescu-Niculescu-Mizil et al., 2011). Register is a strong and important aspect in this, as proposed by Pennebaker, who claims that words can be a “window to the soul” (2011). The linguistic characters that are used can be seen as markers within distinct groups as shown on Wikipedia (Danescu-Niculescu-Mizil et al. 2012). Language coordination is strongly dependent on the accommodation theory and power differences within social groups. On Twitter this is done by posting, replying and retweeting. Their interactions rely on linguistic style markers, such as the use of content and function words and certain keywords (Anger, 2011). All in all, not only the network a person moves in, but also the language is important for analysing online social formations on social media platforms.

The data

Data for this study consist of Dutch Twitter messages from January 2015 to October 2015: in total 561.179 tweets, of which 363.079 were unique tweets and 198.100 retweets. Selected was based on two relevant terms in the refugee debate, both subject to different forms of connotation and representation: “vluchteling” (refugee) and “gelukszoeker” (literally happiness seeker, or fortune seeker or economic refugee). For the purpose of building a corpus for semantic mining the completeness of the data set was of lesser importance than its representation of distinct clusters.

Findings

Our findings address two aspects in political debates on Twitter:

Structure of participants and debate: network analysis confirmed the polarisation within the refugee debate into two opposing clusters, the dynamic of media outlets and opinion leaders in shaping the debate and the interaction of the various participants.

Semantic analysis: a quantitative analysis of language use revealed significant distinctions between the two opposing groups: e.g. the right wing uses more adjectives, needs more words to convey a message and uses more six-letter words than the opposing cluster.

This paper focuses on the distinctions in language use which opens new possibilities for automatically mining Twitter or any corpus for meaning. With regard to recent efforts to create models and ways of algorithmic analysis of social media content (Burnap and Williams 2015), our paper indicates the possibility to move from network analysis to semantic analysis of large corpora. While Ranganath et al. propose a model for predicting protest tweets, our concept suggests the detection of extreme political positions in social media debates (2016). The limitation in Ranganath et al is their dependence on the network structure and history of social interaction of the various participants. In our example, the network provides merely the baseline for further linguistic analysis. Developing this further would entail creating an algorithm based on the findings from our initial corpus. However, this raises issues about the quality of public political debate, freedom of expression and privacy. Data retention and social media metrics provide the means for a coherent analysis of political expressions and deliver powerful tools for security authorities to monitor the political expression online.

References

- Anger, I., & Kittl, C. (2011). Measuring influence on Twitter. Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies.
- Burnap, P. Williams, M. (2015) Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet* 7(2), 223–242.
- Conover, M.D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., Menczer, F. (2011). Political Polarization on Twitter. Proceedings 5th International AAAI Conference on Weblogs and Social Media.
- Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. Proceedings of the 21st International Conference on World Wide Web.
- Danescu-Niculescu-Mizil, Cristian, Gamon, Micheal, Dumais, S. (2011). Mark my words! Linguistic Style Accommodation in Social Media. *WWW*, 78(11).
- Gilbert, E. (2012). Predicting tie strength in a new medium. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work.
- Groh, G., & Hauffa, J. (2011). Characterizing Social Relations Via NLP-based Sentiment Analysis. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 502–505.
- Paßmann, J., Boeschoten, T., & Schäfer, M. T. (2014). The Gift of the Gab: Retweet Cartels and Gift Economies on Twitter. *Twitter and Society*, 331–344.
- Pennebaker, J. W. (2011). Your use of pronouns reveals your personality. *Harvard Business Review*, 89 (December), 32–3.
- Ranganath, S., Morstatter, F., Hu, X., Tang, J., Wang, S., Liu, H. (2016). Predicting Online Protest Participation of Social Media Users. *Association for the Advancement of Artificial Intelligence*.

3. Learning complementary alternative medicine socially? Topic modeling health consciousness with big online discussion forum data

Marjoriikka Ylisiurua, Consumer Society Research Centre, University of Helsinki

Introduction

Individuals have varying abilities to understand and retain health-related information. This *health literacy* forms the basis of individuals' knowledge in making the decisions concerning their health (Chinn, 2011; Sorensen et al., 2012; Walsh & Elhadad, 2014). The concept of health literacy is widely employed in healthcare research as a measurable, rational skill.

Nevertheless, individuals with health literacy levels that standard instruments score as "adequate", often rely on biomedically controversial Complementary and Alternative Medicine (CAM) treatments (Bains & Egede, 2011; Stoneman, Sturgis, & Allum, 2012). One potential explanation comes from Chinn (Chinn, 2011) and Puuronen (Puuronen, 2015, in Finnish), both of whom highlight the effects of social community on what information individuals accept as relevant. Puuronen (2015) refers to "extended" health literacy as *health consciousness*, which includes (sub-) cultural codes and socially constructed meanings. To complement earlier research, this paper analyses a large online data set to study how online communities shape health consciousness in the field of CAM.

The material for this research consists of discussions on the CAM field of homeopathy. Online discussion forums spread "traditional biomedical" knowledge, health experiences and peer advice, making social media a field where health literacy is acquired and required (Centola, 2013; Cline & Haynes, 2001). The task is to analyze discussion topics (DiMaggio, Nag, & Blei, 2013) and to investigate how writers learn health consciousness socially, while expressing health literacy capabilities. To that end, the study employed both the topic modeling algorithm LDA (Blei, Ng, & Jordan, 2003), and close reading.

Materials & methods

Suomi24.fi ("Finland24.fi") is the largest and one of the oldest Finnish discussion forums in which readers and contributors may either register or use a temporary alias. Various discussion subfora consist of discussion threads that engage contributors in conversations which occasionally last as long as several years.

A data set covering Suomi24-activity from years 2001 to 2015 is available for academic use at The Finnish Language Bank Fin-Clarín database⁴⁸. It consists of over 55 million discussion comments and their metadata, e.g. time stamps and contributor nicknames. From this database, the full Homeopathy subforum data set was acquired in CSV format. Each row included an original discussion sentence, its lemmatized sentence with its stopwords removed, and sentence metadata. The data file totals 26MB (52,729 sentences, or 9,326 comments).

As the first step, an LDA algorithm developed with Python Gensim package was run with a varying number of topics. After algorithmic modeling, comments and their surrounding discussion threads

⁴⁸ Resource description: <http://urn.fi/urn:nbn:fi:lb-2017021503>

were analyzed with close reading. The original texts were then sampled using topic model keywords as CSV/Excel search keywords. This resulted in a 15-topic model, combined into frames as follows:

- Fields of controversy: Both “proponents” and “opponents” of homeopathy discussing scientific evidence for/against homeopathy and its areas of application. (topics #A)
- Historical context: Mostly proponents describing the lengths of their personal experiences, as well as the history of homeopathy. (2 topics)
- Celebrity discussion: Mostly proponents discussing a physician who is a public proponent of homeopathy. (1 topic)
- Help gained from homeopathy: Contributors employed this frame to describe their experiences with homeopathy. (topic #B)
- Asking questions: Primarily “new initiates” describing their condition and asking for help. (topic #C).

After recognizing the discussion frames algorithmically, analysis continued with further close reading, including sampled, original online conversations.

Results

Typically, homeopathy “proponents” see the treatment as complementary to traditional biomedicine. The proponents’ positioning of homeopathy in relation to biomedical medicine is revealed in the frames on homeopathic history and celebrities. In this paper however, the focus is on the three remaining frames (#A, #B, #C) to observe health consciousness and health literacy expressions in dialogues between proponents and opponents of homeopathy.

Self-professed “initiates” to homeopathy (topic #C) often seek peer experiences on how homeopathy handles certain conditions. In response (topic #B), some proponents underline the importance of finding the right homeopathic practitioner. In contrast, some proponents describe the use of homeopathic products that they administer independently, without the homeopaths’ advice. Furthermore, some authors describe homeopathy’s failure to cure their condition, whereas “opponents” promote reliance on biomedicine.

Experience and opinion sharing elicits dialogue, which often turns heated. Especially the self-professed, experienced proponents sought to actively defend their individual experiences against attacks from opponents of homeopathy (topic #B), as highlighted below.

Personal experience with my child’s chronic otitis: The child couldn’t stomach many antibiotics due to allergies and a sensitivity to medicines. Medicines would often cause severe symptoms so alternatives were needed, and homeopathy provided the answer. **How could a small child under 3 years old pretend that a substance is helpful, they wouldn’t know.** And yet, homeopathy was the only effective measure. The later check-ups confirmed that the infection was remedied. Our pets, too, have had success with homeopathy, **and they can’t pretend either. A good homeopath can choose a suitable substance. It needs to be the correct one to be effective.**⁴⁹

The homeopathic community supports its proponents facing opponent scrutiny. Some strategies involve capabilities and lexis that suggest high health literacy. For example, opponents may defend biomedical treatments or accuse homeopathy of lacking scientific evidence, using biomedical terms like “Pediatric Neurotransmitter Disorders”⁵⁰. The proponents then express their capability in

⁴⁹ Author “Arnica D”, 31.8.2014, <http://keskustelu.suomi24.fi/t/12228806/homeopatiaa-kokeilleiden-kokemuksia-kaivataan!>

⁵⁰ <http://keskustelu.suomi24.fi/t/2428127/tarkkaavaisuushairio>

counterattacks, citing commercial nature of big pharma or downsides of excessive use of antibiotics, using biomedical terms like “Hospital-Acquired Infections” (topics #A).

This study suggests an inter-group conflict mechanism for social health consciousness learning. The Suomi24 homeopathic community exhibits traces of health literacy, yet positions CAM treatments differently than its opponents. The material and methods obviously do not allow concluding that all comments reflect the authors’ actual experiences and opinions. However, forum discussions may be an important learning environment for contributors and non-contributing visitors alike. To understand the social process of health consciousness in the field of homeopathy, this study should be complemented with interviews and ethnographic research.

REFERENCES

- Bains, S. S., & Egede, L. E. (2011). Association of Health Literacy with Complementary and Alternative Medicine Use : A Cross-Sectional Study in Adult Primary Care Patients. *BMC Complementary and Alternative Medicine*, 11(138), 7. <http://doi.org/10.1186/1472-6882-11-138>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Centola, D. (2013). Social media and the science of health behavior. *Circulation*, 127(21), 2135–2144. <http://doi.org/10.1161/CIRCULATIONAHA.112.101816>
- Chinn, D. (2011). Critical health literacy: a review and critical analysis. *Social Science & Medicine*, 73(1), 60–67. <http://doi.org/10.1016/j.socscimed.2011.04.004>
- Cline, R., & Haynes, K. (2001). Consumer health information seeking on the Internet: the state of the art. *Health Education Research*, 16(6), 671–92.
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606. <http://doi.org/10.1016/j.poetic.2013.08.004>
- LexisNexis. (2007). How Many Pages in a Gigabyte ? Retrieved from https://www.lexisnexis.com/applieddiscovery/lawlibrary/whitePapers/ADI_FS_PagesInAGigabyte.pdf
- Puuronen, A. (Ed.). (2015). *Terveystaju Nuoret politiikka ja käytäntö*. Helsinki, Finland: Nuorisotutkimusverkosto.
- Sorensen, K., Van Den Broucke, S., Fullam, J., Doyle, G., Pelikan, J., Slonska, Z., ... European Health Literacy Project Consortium (HLS-EU). (2012). Health literacy and public health: a systematic review and integration of definitions and models. *BMC Public Health*, 12(1), 80. <http://doi.org/10.1186/1471-2458-12-80>
- Stoneman, P., Sturgis, P., & Allum, N. (2012). Understanding support for complementary and alternative medicine in general populations : Use and perceived efficacy. *Health*, 17(5), 512–529. <http://doi.org/10.1177/1363459312465973>
- Walsh, C., & Elhadad, N. (2014). Modeling Clinical Context: Rediscovering the Social History and Evaluating Language from the Clinic to the Wards., 224–231.
-

4. Web data extraction allows independent evaluation of Global Absolute Poverty⁵¹

Michail Moatsos
Utrecht University

The widely applied “dollar-a-day” methodology identifies global absolute poverty as declining precipitously since the early 80's throughout the developing world. The methodological underpinnings of the dollar-a-day approach have been questioned in terms of adequately representing equivalent welfare conditions in different countries and years [Reddy and Pogge, 2010; Deaton, 2010; Srinivasan, 2010; Aten and Heston, 2010; Sub-ramanian, 2015; Moatsos, 2015]. If empirically substantiated, such criticism directly questions the validity of the dollar-a-day methodology since in international poverty measurement “the first-order issue is to demand welfare consistency” [Ravallion, 2015, p.4].

However, an independent examination of the levels and trends of global poverty is a very demanding task. In most of its part this is due to the restricted access on national economic distributions of income or consumption that are utmost essential for the calculations. The easiest way to use those distributions is the PovcalNet website offered by the World Bank. Unfortunately, the Bank does not make the underlying distributional data available. Instead it only conditionally allows direct calculations of global poverty. The condition being that the independent researcher accepts the validity of the dollar-a-day approach that the World Bank follows rather religiously. Thus a serious problem appears when one seeks to evaluate global poverty using a different methodology, or when one tries to quantify the aforementioned concerns against the World Bank methodology.

The solution to this conundrum is the use of IT methods that scrap the underlying data from the PovcalNet website, so that they can be used independently of any conditions. This has been done in a first step by Dykstra et al. [2014], but with several issues of discrepancy between the data offered by PovcalNet and those made available by the authors. Based on their work I simplify the process one step further by allowing automated evaluation of poverty in bulk based on independently calculated poverty lines. This approach has the advantage of querying the PovcalNet service “as it is” without discrepancies.

The alternative to the dollar-a-day methodological approach is to estimate absolute poverty on a global level using appropriately defined consumption baskets for each country and year separately. Allen [2001] defines the BBBs for use in the historical real wages literature, and de Zwart et al. [2014] apply them on a global scale. Table 1 contains the overview. The BBBs are constructed such as to represent bare minimum absolute poverty levels in consumption terms. However, the absolute poverty yardstick can be expanded to account for other essential elements of life and wellbeing, such as education and health, as both the Copenhagen Declaration and the Universal Declaration of Human Rights stipulate. Table describes one such BBB derivative that allows for considerably higher welfare levels compared to the basic BBB.

⁵¹ This paper is largely based on a forthcoming article in the Journal of Globalization and Development entitled “Global Absolute Poverty: Behind the Veil of Dollars”.

Table 1: The composition of bare bones baskets in real wages and the two derivatives applied here.

Item	Unit/Year	Real Wages Basket	BBB	BCS
Energy Target	kcal		MDER	MDER
		1455/2100		
Minimization	-	cheapest bundle		mean of 3 cheapest bundles
Main staple	kg		based on kcal/protein target**	
		155-413*		
Beans or peas	kg	-/20/45	LP	40 at minimum
Meat or sh	kg	3 or 6	3 or 6	12 or 24
Butter or oil or ghee	kg	3	3	12
Sugar	kg	-/2	2	8
Linen (applied)	share	8%	8% ± 2%	WBGC
Lamp oil	liter	1.3	1.3	WBGC
Soap	kg	1.3	1.3	WBGC
Candles	kg	1.3	1.3	WBGC
Fuel	mbtu	3	f(T in °C)	WBGC
Cooking	mbtu	-	MDER	WBGC
Housing	mark-up	5%	5% ± 2%	WBGC
Health, Education, Water	%	-	-	WBGC
Additional shares	%	-	-	WBGC

Note: The Bare bones basket with Consumption Shares (dubbed BCS) uses the average of three cheapest bundles, and four times more meat/fish, butter and sugar allowance. In addition, an allowance covering health, education, and water is included using the consumption budget shares from the World Bank Global Consumption dataset (noted as WBGC on the table). Consumption budget shares are also used for energy, housing, and clothing, and allowances for Personal Care, ICT, Financial Services, and Others are included in the additional shares.

*: depending on the country and main staple

** : To avoid inflating the price of the consumption bundle, priority in linear programming is given to the kcal target, and protein target is allowed to overshoot by 200% at maximum if necessary. Only for Dominican Republic this cap increases the bundle price by more than 20%, and for Belarus by more than 10%, compared to allowing for unlimited protein overshooting. For all other countries there increase if any is restricted to only a few percentage points increase.

The results (figure 1) show that, in terms of levels, on the one hand the target of alleviating absolute poverty is not as far off as was thought of, but on the other hand,

absolute BBB poverty has shown remarkable persistence throughout the period. The difference with the PovcalNet estimates is enormous throughout. Comparing the 1990 and 2014 estimates leaves little room for celebrations over the achievement of halving absolute global poverty between 1990 and 2015⁵². Using the BBB poverty lines the point estimate for global poverty in 1990 is 5.6% and for 2014 3%. Using the BCS poverty lines, the corresponding rates are 62% and 33%. In turn, this shows that the conclusion about the questionable MDG1 success does not result from the very low welfare level that the BBB poverty lines encapsulate.

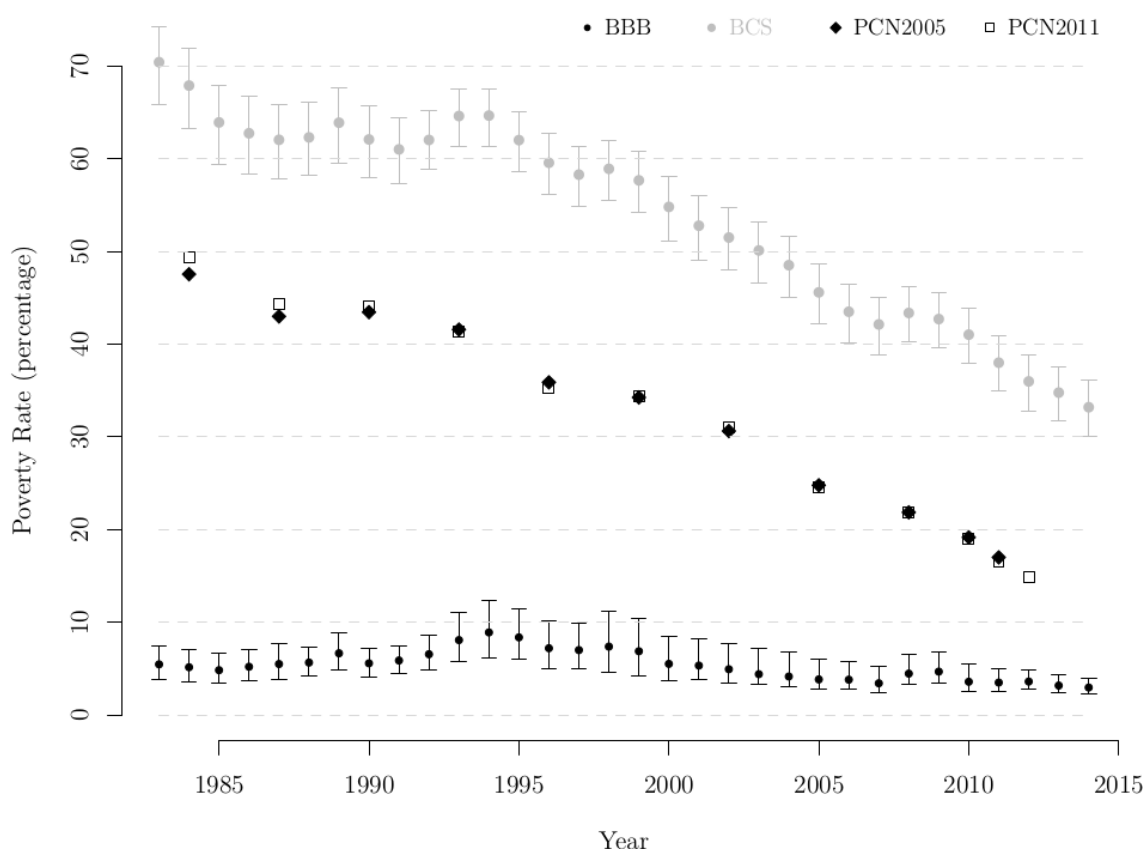


Figure 1: Evolution of poverty in the Developing World, 1983-2014. PCN2005/11 refer to the World Bank global poverty estimates based on the 2005 or 2011--so called--ICP rounds.

The vast differences among BBB welfare level and the International Poverty Line (iPL) can be attributed on two elements. First, the much lower costs of bare bones subsistence compared to the \$1.9 value for the vast majority of the countries and years. And second, on the differential between consumer price index and the BBB price index. The also very large differences of iPL with BCS, especially on the later years of the period, is attributable to the inability of the iPL to encapsulate expenses that are necessary in escaping absolute poverty as described in international treaties and conventions.

This research demonstrates that the use of digital techniques for scraping online data that are not explicitly provided for downloading can provide answers to big questions such as the level and trends of global poverty. It is important that this work can be performed independently of the custodian

52 Millennium Development Goal 1: "Target 1.A: Halve, between 1990 and 2015, the proportion of people whose income is less than \$1.25 a day". The World Bank has announced that this goal has been achieved as early as 2010, five years ahead of schedule.

institutions for monitoring and reducing poverty (the World Bank). Institutionally based decisions, evaluations and calculations are not necessarily beyond dispute; and must not be. The next step in this project is to elaborate on the proper accounting of uncertainties in the estimates using the Monte Carlo method for pseudo-experiments. This, computationally very demanding task, would allow for a more appropriate comparison of the poverty estimates between the target years of MDG1.

References

Allen, R. C. (2001). The Great Divergence in European Wages and Prices from the Middle Ages to the First World War. *Explorations in Economic History*, 38:411-447.

Aten, B. and Heston, A. (2010). Use of Country Purchasing Power Parities for International Comparisons of Poverty Levels: Potential and Limitations. In Anand, S., Segal, P., and Stiglitz, J. E., editors, *Debates on the Measurement of Global Poverty*. Oxford University Press, Oxford.

de Zwart, P., van Leeuwen, B., and van Leeuwen-Li, J. (2014). Real Wages. In van Zanden, J. L., Baten, J., D'Ercole, M. M., Rijpma, A., and Timmer, M. P., editors, *How Was Life? Global Well-being since 1820*, chapter 4, pages 73-86. OECD Publishing, Paris.

Deaton, A. (2010). Measuring Poverty in a Growing World (or Measuring Growth in a Poor World). In Anand, S., Segal, P., and Stiglitz, J. E., editors, *Debates on the Measurement of Global Poverty*, pages 187-222. Oxford University Press.

Dykstra, S., Dykstra, B., and Sandefur, J. (2014). We Just Ran Twenty- Three Million Queries of the World Bank's Website.

Moatsos, M. (2015). Global Absolute Poverty: Behind the Veil of Dollars. CGEH Working Paper Series, (77).

Ravallion, M. (2015). Toward Better Global Poverty Measures. Center for Global Development, Working Paper (417).

Reddy, S. G. and Pogge, T. (2010). How not to count the poor. In Anand, S., Segal, P., and Stiglitz, J. E., editors, *Debates on the Measurement of Global Poverty*, pages 42-51. Oxford University Press.

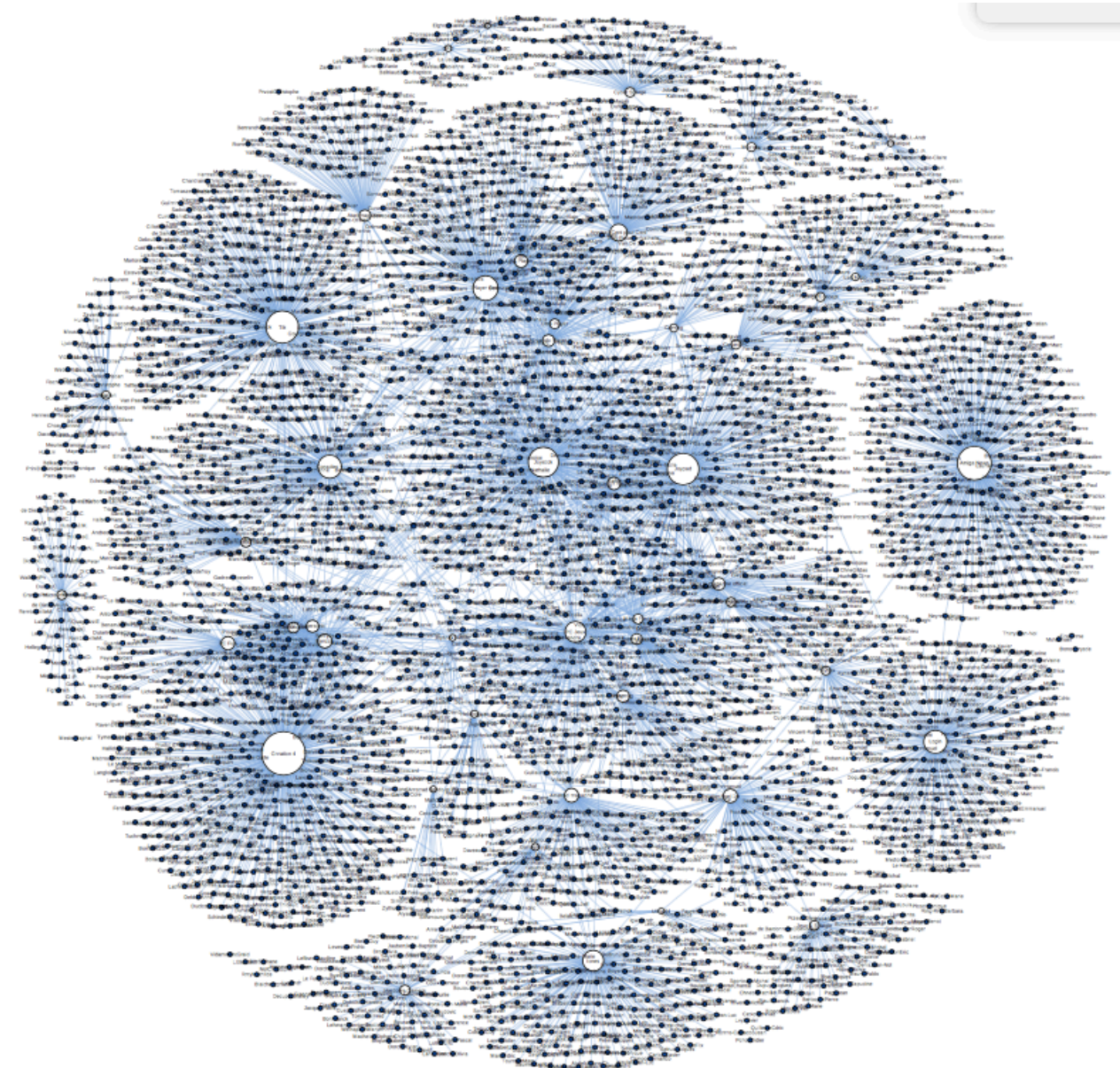
Srinivasan, T. N. (2010). Irrelevance of the \$1aDay Poverty Line. In Anand, S., Segal, P., and Stiglitz, J. E., editors, *Debates on the Measurement of Global Poverty*, pages 143-151. Oxford University Press, New York.

Subramanian, S. (2015). Once more unto the breach. *Economic & Political Weekly*, L(45):35-40.

1. Sociology of French Video Game Magazines

Björn-Olav Dozo

The first video game magazine in French, *Tilt*, was published by Editions Mondiales in September 1982, just a few months after the first release of *Computer and Video Games* (UK, November 1981) and *Electronic Games* (US, November 1981). It established a model for future French-speaking video game magazines, with a stable structure (news, previews, tests) present in any magazine until the early 2000's.



Graph of the relations between magazines of the corpus.

The 1990's are a very profitable decade for these magazines as the editorial field is structured to support game developers, with a pro-Nintendo pole and a pro-Sega pole. While magazine titles stood in rhetoric opposition (*Super Power* pro-Nintendo vs. *Mega Force* pro-Sega, about 120 000 monthly copies each), they shared the same editorial boards: the same journalists wrote in different

magazines of one publisher, but with different pseudonyms. At times, they simulated competition between the various editorial boards, giving to the readers the feeling of belonging to a community. This kind of strategies was common until 1996, but when a new challenger (Sony) came into the dance, some magazines chose to merge with old competitors of the same press group in order to survive.

In 2003, “Future France” bought almost all the video games magazines titles available on the French market. This hegemonic strategy, however, has not proven to be profitable on the long term: a lot of these titles, even long-running magazines with faithful audiences, discontinued their publication in the years following the buyout. My talk will question the context of these cessations of activities. Different reasons could be given: the internet explosion of video games information’s websites, the weakness of the economic model of the paper press or the demotivation of journalists. Other initiatives emerged at this time, as Canard PC and Gaming for example, proposing a different business model (independent press). After this first stage, I will further analyse the career-path of these specialized journalists with a social network analysis, following their path between different redactions in this very small world. The database that I use is compiled from the examination of about 80 titles of French-speaking video game magazines over 30 years. With these data, I will show the evolution of the field, with the migration of some journalists between different publications, sometimes on the basis of a kind of “mercato” of local writing stars.

Bibliography

Bae, Arram, Doheum Park, Yong-Yeol Ahn, and Juyong Park. 2016. ‘The Multi-Scale Network Landscape of Collaboration’. *PLOS ONE* 11 (3): e0151784. doi:10.1371/journal.pone.0151784.

Falk, Casey. 2014. ‘Using Network Analysis on the Du Chemin Music Dataset to Reconstruct Missing Music’ [unpublished paper].

Giannetti, Francesca. 2016. ‘A Review of Network Approaches in Music Studies’. *Music Reference Services Quarterly* 19 (2): 156–63. doi:10.1080/10588167.2016.1166842.

Gresham-Lancaster, Scot. 2014. ‘Computer Music Network’. *Leonardo* 47 (3): 266–67. doi:10.1162/LEON_a_00771

2. Musical networks – Networks of music

**Marnix van Berchum,
Utrecht University**

Network models are widely used in Digital Humanities for understanding relational structures in information. The available mathematical tools used in network science allow scholars to analyse their material in a quantitative manner, and for example find relative centrality measures for certain network entities or discover community structures. Visualisation tools assist in discovering large scale patterns in the network, pointing to areas where a more thorough, qualitative analysis is needed. The presence of network related contributions in the programmes of Digital Humanities conferences, the ongoing emergence of new tools for building and visualising networks, and the many humanities projects making use of these publications and tools attest to this popularity. With a different level of intensity, most Humanities disciplines make use of network methodologies. There is for example a strong community of historians working with/on networks, demonstrated by the extended bibliography and overview of tools at <http://historicalnetworkresearch.org>. The sessions of the ‘Arts, Humanities and Complex Networks’ satellites at the yearly *NetSci* meetings show a variety of disciplines – including art history, film history, literary history and musicology – making use of the methods and tools of network science.

In recent years a growing number of publications appeared, that combines networks and music. The subjects are wide ranging, from social networks between seventeenth composers (Smith and Taylor 2014), to co-occurrence networks of composers on CD recordings (Park et al. 2015; Bae et al. 2016), to “Computer Music Network” of the late 70s / early 80s (Gresham-Lancaster 2014). Similar to the range of subjects a variety of network methodologies applied to music is discernible in these publications. I will compare the approaches of the selected publications and answer questions on how they relate to the more ‘traditional’ musicological discourse. The paper will discuss the biases present in the data used in the publications and how these effect the musicological conclusions made. It touches upon the tension between the quantitative (‘distant’) character of network science and the qualitative (‘close’) character of musicology research.

The aim of my own PhD-research is to bridge this gap. In my research I use a network approach to shed light on the dissemination of polyphonic music in the sixteenth century, the age of the emergence of printed music. Primary musical sources and the compositions they contain are the two entities that form the network. Since one source contains several compositions, and one specific composition may be present in multiple sources, a bipartite network of sources and compositions comes into existence. Both extensive musicological studies of these sources and compositions, as well as high level network structures are used to formulate a model for the dissemination of music. In this paper I will compare and relate my own experience with the evaluation of the selected publications, concluding with an insight into what networks, social network analysis and related methods and tools offer – and may offer in the future – to the field of Musicology.

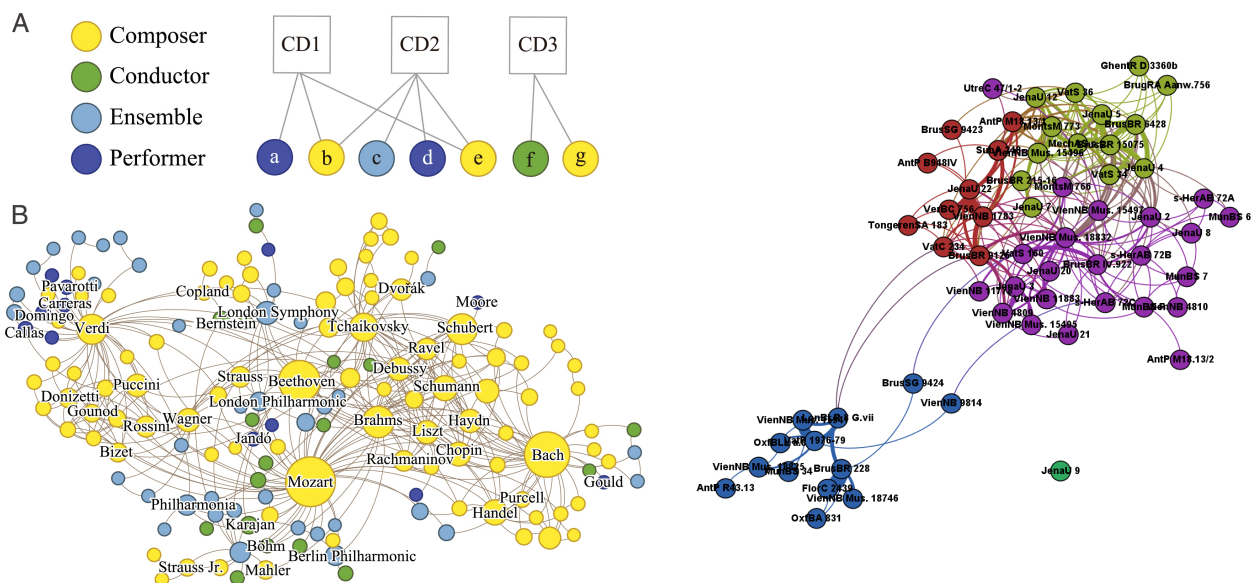


Illustration taken from Bae et al. 2016, showing the co-occurrence of musicians (composers and performing artists) on CD-recordings; data taken from ArkivMusic.

Illustration from the author’s PhD research, showing the network of manuscripts from the Alamire scriptorium; each node represents a manuscript, an edge represents at least one musical composition two manuscripts have in common.

Bibliography

Bae, Arram, Doheum Park, Yong-Yeol Ahn, and Juyong Park. 2016. ‘The Multi-Scale Network Landscape of Collaboration’. *PLOS ONE* 11 (3): e0151784. doi:10.1371/journal.pone.0151784.

Falk, Casey. 2014. 'Using Network Analysis on the Du Chemin Music Dataset to Reconstruct Missing Music' [unpublished paper].

Giannetti, Francesca. 2016. 'A Review of Network Approaches in Music Studies'. *Music Reference Services Quarterly* 19 (2): 156–63. doi:10.1080/10588167.2016.1166842.

Gresham-Lancaster, Scot. 2014. 'Computer Music Network'. *Leonardo* 47 (3): 266–67. doi:10.1162/LEON_a_00771.

Park, Doheum, Arram Bae, Maximilian Schich, and Juyong Park. 2015. 'Topology and Evolution of the Network of Western Classical Music Composers'. *EPI Data Science* 4 (1). doi:10.1140/epjds/s13688-015-0039-z.

Piekut, Benjamin. 2014. 'Actor-Networks in Music History: Clarifications and Critiques'. *Twentieth-Century Music* 11 (02): 191–215. doi:10.1017/S147857221400005X.

Smith, David J, and Rachelle Taylor. 2014. *Networks of Music and Culture in the Late Sixteenth and Early Seventeenth Centuries: A Collection of Essays in Celebration of Peter Philips's 450th Anniversary*.

3. The “Frame Generator”. An alternative method for approximating core meanings in texts

Joris van Eijnatten (Universiteit Utrecht)

Juliette Lonij (Koninklijke Bibliotheek)

Tracing semantic patterns over time on the basis of texts is still in its infancy. Most approaches build on a linguistic principle which states that the meanings of words are determined 'by the company they keep'. In other words, meanings arise from contexts defined as distributions of words, which suggests that we can trace meanings over time by examining changing contexts. Topic modelling is at this moment the only technique based on the principle of word distributions that has gone beyond an experimental stage and has proven its value by achieving results that domain experts (in this case historians not necessarily involved in computer-assisted research) recognize.

This paper discusses a new tool, dubbed the 'Frame Generator', aimed at meaningfully reducing a set of (possibly thousands of) Dutch texts to word patterns that cut across the distributions generated by topic modelling, thus providing additional insight into the content of the dataset. The method implemented builds on topic modelling by combining it with two other proven techniques: (1) the automatic extraction of keywords and (2) the identification of collocates. The Python source code of the tool, offering a command line interface, is available for download on GitHub (<https://github.com/jlonij/frame-generator>). An online demo with a graphical user interface, showcasing the tool's main functionality for a small dataset, can be found at <http://kbresearch.nl/frames/>.

The Frame Generator was developed to assist in the investigation of popular perspectives on the concept of 'Europe' arising from the KB collection of Dutch historical newspapers. To this end, a dataset was prepared of articles that mentioned the word 'Europe' at least once. A subset of articles was then selected on the basis of (Dutch-language) synonyms for the words 'unity' and 'unification' (such as 'integration', 'agreement', 'settlement', 'consensus', 'treaty', 'harmony', etc). This subset was assumed to contain news articles that discuss Europe as a unified political / cultural / economic entity, or as an entity involved in a process of unification. The other subset was based on synonyms for competitions (such as 'match', 'prize', 'winner', 'cup', etc); this subset was assumed to contain articles on sports and other competitions.

The Frame Generator process of analyzing these datasets consists of four stages. The first stage concerns the pre-processing of the dataset. During this stage the dataset is cleaned by normalizing spelling variations and correcting OCR errors on the basis of user-provided lists of regular expressions and their replacements. In addition, the dataset is tokenized, lemmatized and part-of-speech tagged with the Natural Language Processing suite Frog (<https://languagemachines.github.io/frog/>). The user has the option of splitting larger documents into smaller units of analysis by specifying the maximum number of sentences to be contained in each unit.

The second stage in the process is topic modelling, which generates specific, substantive themes or topics based on frequently recurring distributions of words. The Frame Generator offers two methods of topic modelling: one based on Mallet (<http://mallet.cs.umass.edu>), the other on the Gensim topic modelling library). The user is able to control the number of topics generated and number of words making up each topic by means of various command line arguments. This stage also involves the manual, hermeneutic interpretation of the topics based on historical domain knowledge.

The third stage focuses on the extraction of a single, ranked list keywords from the set of topics resulting from the previous stage. The relevance of each word occurring in the set of topics is determined by taking the sum of the probability scores for the word over all topics in which it occurs. A word is accorded the status of keyword if its score reaches a certain threshold, set at the discretion of the researcher. The Frame Generator can also produce a keyword list on the basis of tf-idf scores, thus allowing the researcher to compare the results of different approaches. The option is available to restrict the candidates for the keyword list to words with specific part-of-speech tags. The keywords thus obtained may be regarded as core elements in a series of thematically uniform texts; their significance arises from the frequency of their occurrence within as well as across topics.

The fourth and final stage of the analysis process consists of contextualising the keywords by finding collocates in the texts from which they were originally extracted. The user sets a maximum word distance from the keyword as well as the direction (left, right, or both) in which collocates must occur in order to qualify. As with the extraction of keywords, the option to include only specific part-of-speech tags is also provided for collocates. The set of collocates thus gathered for a given keyword is called a 'frame'. The words appearing in a frame are ordered by the frequency of their co-occurrence with and their distance to the keyword with which they are associated, expressing their significance in framing a specific keyword.

The results of each of these stages are saved and accessible to the user in the form of comma-separated values (CSV) files. These can, for example, be used to visualise the graph of the keywords and their collocates in an application such as Gephi (<https://gephi.org>) in order to facilitate the interpretation of the results. By creating such network graphs for the Frame Generator results for a number of different time periods (see Figure 1 for an example) we found that newspaper reporting on 'European unity', while showing a remarkable degree of continuity, became less rich rhetorically, less international, and more focused on institutional technocracy than on intra-continental relations over the course of the twentieth century.

This paper hypothesises that the Frame Generator, by laying bare the fundamental patterns in sets of thematically coherent texts, enables historians to better determine continuities and discontinuities in expressions of public opinion. The Frame Generator's performance depends on that of its constituent tools (such as topic modelling), which have been described in the literature. Its advantages include its adaptability to other languages (given the availability of part-of-speech tagging), its flexibility (the user can set all variables) and its 'all-in-one' packaging (it requires no programming skills while generating not just frames but also keywords and topics). For domain experts (historians) the proof of the pudding will be in the eating: does this particular combination of tools – topic modelling, keyword extraction and identification of keyword collocates – offer useful results? The question can only be answered by running the tool on a variety of relatively homogenous datasets.

while most historians are accustomed to deploying digital approaches in the information gathering stage of their research, they often refrain from 'going digital' in its processing and especially analysis stages. Describing a number of digital tools used in work done on the diaries of Anne Frank, the paper critically analyses and demonstrates the added value of incorporating them in all stages of historical research. Digital approaches enhance the methodological repertoire furnished by 'traditional' close reading practices. Hybrid approaches thus expand our intellectual horizons and the analytical power we bring to bear upon our sources.

The paper consists of 1) a theoretical part, contextualising notions of 'traditional' and digital approaches to historical research and the uptake of the latter; and 2) a concrete case study of a hybrid approach to historical research.

The first part will briefly discuss discourses around 'going digital' that often oppose 'traditional' to digital approaches. On a general level, this either/or attitude is misleading; despite what is often assumed, or implicitly suggested, distinctions cannot be neatly mapped along lines of close reading/distant reading, quantitative/qualitative or positivist/narrative analysis either. More specifically, this opposition is also problematic because, for instance, close reading can also involve the use of digital tools, and the same obviously goes for qualitative analysis. In this respect, one should mention Frédéric Clavert's use of Franco Moretti's concept of 'distant reading' to propose a new way of reading and interpreting historical sources in the digital age using two axes – close reading/distant reading and human reading/computational reading.

The focus will then shift to the problem of uptake of digital approaches among historians. Here, a distinction is drawn between two broad strands of historical research in the digital era, as measured by their application of digital approaches:

- on the one hand, a number of digital historians take (big) data and digital tools (development) as their point of departure; their focus is on digital datasets (for instance newspapers) and the application of digital tools to an analysis of that data. This yields research results that are often as much, if not primarily, concerned with critical reflection on data and tools as with the research topic at hand. Tool development is often also part of the process and project and research questions tend to be dictated by the available data and tools.
- on the other hand there are those historians, arguably the majority, whose research does not start with data sets and tools; they depart from particular research questions pertinent to their topic of research that could be answered, at least in part, by digital means; the question then becomes how digital approaches can aid, enhance and complement their analyses. As will be clear, the problem of uptake is pertinent to this group of historians.

The second part of the paper concerns itself with a hybrid analysis of a concrete historical source: the diaries of Anne Frank. This part of the paper is one of the outcomes of a three-year research project (The diaries of Anne Frank. Research—Translations—Critical Edition) which was carried out at the Lichtenberg Kolleg of the Georg-August-Universität Göttingen. The project involves a new scholarly edition of the diaries as well as an accompanying multi-author research monograph which will focus on contextualisation, reception and representations of the diaries. Describing a number of digital tools (notably text mining and QDA software) used in analysing the diaries, the paper critically analyses and demonstrates the added value of incorporating them in all stages of historical research. The aim here is to apply Clavert's basic model, as mentioned above, to a concrete case study and, ultimately, to provide historians with a concrete example of a hybrid approach to historical research.

1. Towards a tool and data criticism framework

A developer's and user's perspective

Sally Chambers¹, Joke Daems¹, Greta Franzini², Marco Büchler², Susan Aasman³

¹Ghent Centre for Digital Humanities, Ghent University

²Institute of Computer Science, University of Göttingen

³Groningen Centre for Digital Humanities, University of Groningen

As the amount of accessible digitised data grows, so does the need for machine assistance to help process this overload of information. As cultural heritage institutions increasingly digitise their collections, they are in effect converting the collections into data. Particularly in the area of Digital Humanities, the need for 'full-text' collections for analysis, is becoming increasingly important. For example, in 2016 the National Library of the Netherlands organised a workshop 'Historical Newspapers as Big Data'.⁵³ The focus of this workshop was to bring together researchers from a range of disciplines who were interested in using the digitised newspapers and other digital collections made available by the *Delpher platform*⁵⁴ for (digital) humanities research.

In recent years, international initiatives such as the *DiRT Digital Research Tools* directory⁵⁵, the *Common Language Resources and Technology Infrastructure* (CLARIN)⁵⁶ and the *Digital Research Infrastructure for the Arts and the Humanities* (DARIAH)⁵⁷ have been bringing together tools and resources to help scholars repurpose data for the advancement of research and knowledge.

Despite the proliferation of tools, little is known about their development and use. As Gibbs and Owens observed (2012), this grey area of knowledge concerns both the production and the user side of tools, raising questions about usability, purpose, effectiveness and usage.

From a development standpoint, often assumptions are made with regard to users and the use of a tool. While tools are typically designed to be part of the solution to a problem, by assuming knowledge they become part of the problem to be solved.

From a user perspective, perhaps the biggest barrier to the adoption of a tool is the absence of (sufficient) documentation on their application (i.e. "how to" instructions) and on their functionality (i.e. the "black box"). Functionality is key to the evaluation of computed results, in that if the inner workings of a tool are opaque, how or to which extent can the user trust the results? How useful is the tool?

Whereas developers need to be clear about what the tool is intended for, users need to be careful in selecting the appropriate tool to address their research question. An improved understanding of both the developer's intentions in tool development as well as the user's requirements in order to answer their research question are needed. Additionally, the particular data-set that a user wishes to analyse, is a crucial factor when it comes to tool selection.

⁵³ See: <https://www.kb.nl/nieuws/2016/historische-kranten-als-big-data-ii-concepten-op-drift> (Accessed: 9 February 2017).

⁵⁴ Available at: <http://www.delpher.nl/> (Accessed: 9 February 2017).

⁵⁵ Available at: <http://dirtdirectory.org/> (Accessed: 9 February 2017).

⁵⁶ Available at: <https://www.clarin.eu/> (Accessed: 9 February 2017).

⁵⁷ Available at: <http://www.dariah.eu/> (Accessed: 9 February 2017).

In light of the issues described, this contribution reiterates the need for tool criticism, previously expressed at the ‘Tool Criticism for Digital Humanities’ workshop (Traub and Ossenbruggen, 2015).⁵⁸ We argue for tool criticism as a pedagogical and effective means of tackling the interdisciplinary challenges posed by the Digital Humanities and of fostering communication between developers and users. Furthermore, we propose an extension of the tool criticism framework to also include data. As an integral part of the research process, we argue that the data-set is an important factor to consider when selecting the appropriate research tools. For example, if a user has a choice between two tools with equivalent functionality, then the structure of the chosen data-set may perform better with one or other of the tools. Additionally, we propose that ‘data criticism’ is an important element in its own right. For example, it is important to critically select the source of a particular data-set, based on a range of criteria. If a particular text is needed for analysis, it may be available from multiple sources. A framework to facilitate the selection of the most appropriate data source is therefore needed. This will build on existing ‘source criticism’ and ‘information evaluation’ frameworks (Hjørland, Birger, 2012).

As a first step towards tool and data criticism, we propose a number of evaluation criteria that seek to encourage a more critical approach to tools. These build upon analogous software studies (Jackson et al., 2011), the EVALITA campaigns⁵⁹ and the very recent *RIDE Digital Text Collections* evaluation guidelines⁶⁰, and are grouped as follows:

Tools

1. Usability
 - a) User Experience (UX)
 - b) Graphical User Interface (GUI)
2. Documentation
 - a) Provenance (authors / organisations behind the tools)
 - b) “How to instructions”
 - c) Algorithms or methods implemented
 - d) Limitations
 - e) Target audience/research
 - f) Availability of tutorials to train users to proficiently work with the tool
 - g) Access and citation
 - h) Rights
3. Maintenance
 - a) Development responses to user feedback
4. Flexibility/Extent of Applicability

Data-sets

5. (Re-)Usability
 - a) Format(s)
6. Documentation
 - a) Provenance (curators / organisations behind the data-sets)
 - b) Metadata (e.g. size, source, author, etc.)
 - c) Limitations
 - d) Access and citation

⁵⁸ See: <http://event.cwi.nl/toolcriticism/> (Accessed: 12 February 2017).

⁵⁹ For more information about *Evaluation of NLP and Speech Tools for Italian* (EVALITA), see: <http://www.evalita.it/2016> (Accessed: 12 February 2017).

⁶⁰ See: <http://ride.i-d-e.de/reviewers/call-for-reviews/special-issue-text-collections/> (Accessed: 13 February 2017).

- e) Rights
- 7. Maintenance
 - a) Development responses to user feedback

We evaluate our criteria on three different projects - one *data-set project*, one *tool* and one *application* of our selected tool on our selected data-set - to compare the user and developer perspectives.

The intention is to foster an understanding of tool and data criticism towards a dialogue between users and developers, including how such a framework could be put into practice.

References

- Gibbs, F., Owens, T. (2012) 'Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs', *Digital Humanities Quarterly*, 6(2) [Online]. Available at: <http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html> (Accessed: 12 February 2017).
- Hjorland, Birger (2012) 'Methods for evaluating information sources: An annotated catalogue', *Journal of Information Science*, 38.3 (June 2012): 258-268.
- Jackson, M., Crouch, S., Baxter, R. (2011) *Software Evaluation: Criteria-based Assessment* [Online]. Available at: <https://www.software.ac.uk/sites/default/files/SSI-SoftwareEvaluationCriteria.pdf> (Accessed: 12 February 2017).
- Traub, M. C., Ossenbruggen, J. van (2015) *Workshop on Tool Criticism in the Digital Humanities: TechReport* [Online]. Available at: <http://oai.cwi.nl/oai/asset/23500/23500D.pdf> (Accessed: 12 February 2017).

2. Supporting Digital Humanities in Dealing with Quality of Web Documents

Davide Ceolin
Lora Aroyo

Vrije Universiteit Amsterdam
de Boelelaan 1081a
1081HV Amsterdam
The Netherlands

d.ceolin@vu.nl
lora.aroyo@vu.nl

Julia Noordegraaf

University of Amsterdam
Turfdraagsterpad 15
1012XT Amsterdam
The Netherlands

j.j.noordegraaf@uva.nl

This paper discusses the development of a new approach for assessing the quality of online documents, contributing a new methodological reflection on online source criticism. Online documents are, in fact, a useful source of information for very diverse groups of users, ranging from researchers and journalists to government officials, activists or parents. However, this information is only useful if we manage to filter out the spam and, most importantly, if we manage to retrieve the documents that better fit the qualitative requirements that specific users have. For example, while for laymen neutrality and readability may be important, for scholars accuracy and completeness may be more relevant.

Assessing the quality of online documents is a challenging task because of their intrinsic peculiarities: their volume, variety, and velocity make it impossible for humans to process them manually. A

combination of human and automated processing needs to be devised to handle their quality assessment. Moreover, quality assessment is a challenging task on its own. The overall quality of a given document is the result of the aggregation of multiple facets (or quality dimensions), such as accuracy, completeness, and neutrality. How these facets are quantified and aggregated is mostly a subjective and context-dependent matter. Users with different tasks at hand have different qualitative requirements. Also, users with different backgrounds are likely to evaluate the same document in a different manner.

A general definition of quality is ‘fitness for purpose’, whereby ‘fitness’ varies with both context and purpose. Although this means that the assessment of the quality of online documents is a flexible, fluid process, we believe it is not impossible to model it. To do justice to the fact that different purposes imply different qualitative requirements (e.g., for writing a newspaper article, source neutrality may be less relevant than accuracy), it is crucial to create a reference system that allows for the quantification of document qualities (e.g., the extent to which a given document is neutral or accurate). When this reference system exists, then we can identify the most accurate, precise, or neutral documents that correspond to those of higher quality for a given task (see Figure 1).

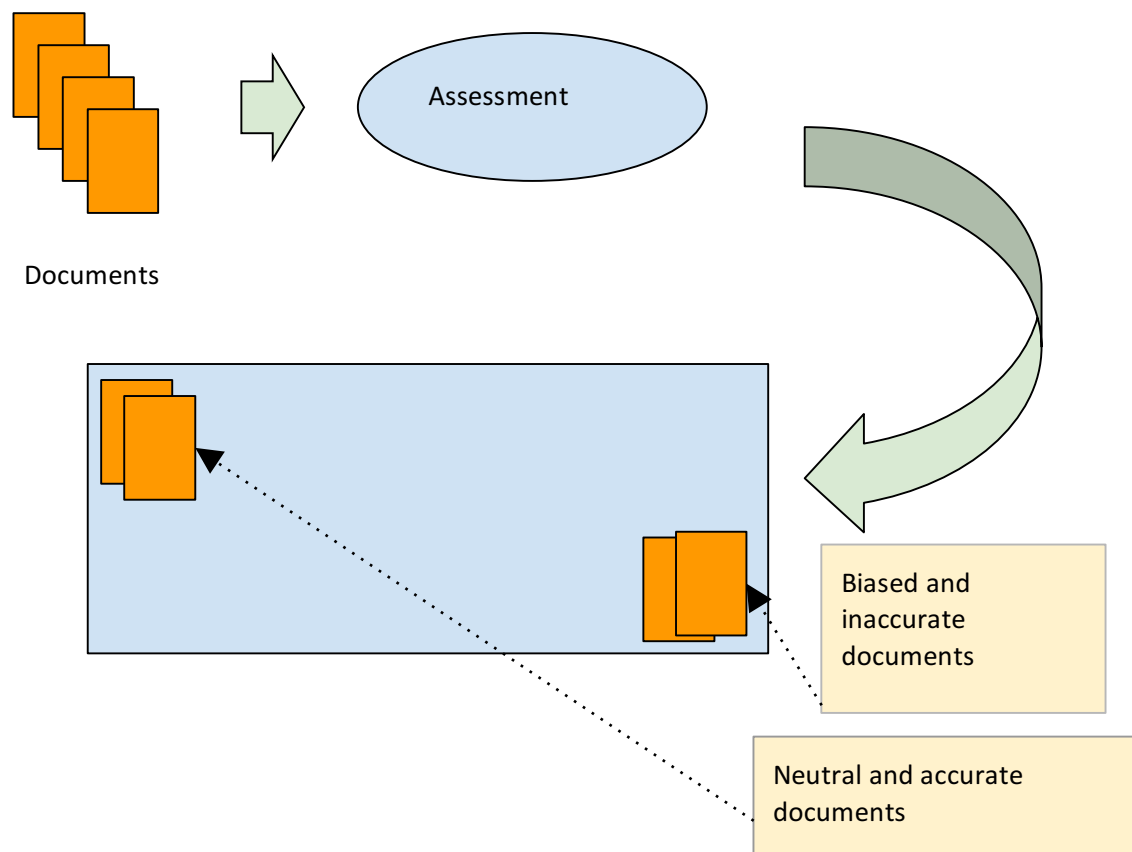


Figure 1. Reference system. Through the assessment, we create a reference system for assessing the relative quality of the documents (e.g., high accuracy, neutrality, or both).

For the purpose of creating a reference system, we are benchmarking a large portion of online documents by employing a combination of human assessment and machine learning. In a pilot study, we used this human-machine interaction approach to assess a selection of online documents [1]. These documents focussed on the topic of vaccinations, and they were selected to provide an overview of the types of documents (blog posts, official documents, etc.), stances (pro, anti, neutral), and types of sources (government authorities, activists, etc.). These documents were assessed by experts (journalists and media scholars) who were asked to judge their relevance for writing an article on the vaccination debate. The experts were asked, first, to judge relevance based on certain automatically generated quality features (such as the trustworthiness of the document, the entities

mentioned and the sentiment expressed in it) and, second, to highlight parts of the document and manually annotate the quality features (such as provenance, references, specific statements in the text itself). The results collected showed that the subjectivity of these assessments is limited by the fact that contributors share a similar background and by the clear definition of the task proposed (documents were assessed supposing they were used for a specific task). This exploratory study provided promising indications for automating and scaling up this process.

Currently, we are employing crowdsourcing to extend the coverage of human assessments of Web documents. By employing the crowd in place of niche experts, we can extend the number of documents assessed. Nevertheless, such a shift requires the document assessment tasks to be simplified because of the different typology of contributors, and because crowdsourcing tasks are usually shorter than nichesourcing tasks [2]: assessing the quality of Web documents is a lengthy and demanding task. However, since the crowdsourcing version of these tasks is intended to capture implicit quality evaluations that users usually do when reading online documents, such a simplification will affect the granularity and not the reliability of the results. For example, we still ask the contributors to assess the precision, completeness, and neutrality of documents, but we use a Boolean scale instead of a Likert scale, and we limit the depth of the argumentations requested.

We are also exploring the possibility to automate such assessment process (see Figure 2). We extracted a set of features from the documents in our pilot study. These included NLP features (e.g., named entities, sentiment analysis) and provenance (e.g., source trustworthiness), and we found that it is possible to employ algorithms like Support Vector Machines [3] to predict the quality assessments by using these features with an accuracy up to 72%. We are extending this prediction, to scale up the number of documents assessed and to improve the accuracy of the predictions. We are scaling up the process of feature extraction, by parallelizing the natural process analysis to extract textual features from large collections of documents. We are also scaling up the prediction part, which takes as input the features extracted and the training data provided by the crowd and by the niches, and produces the quality estimations.

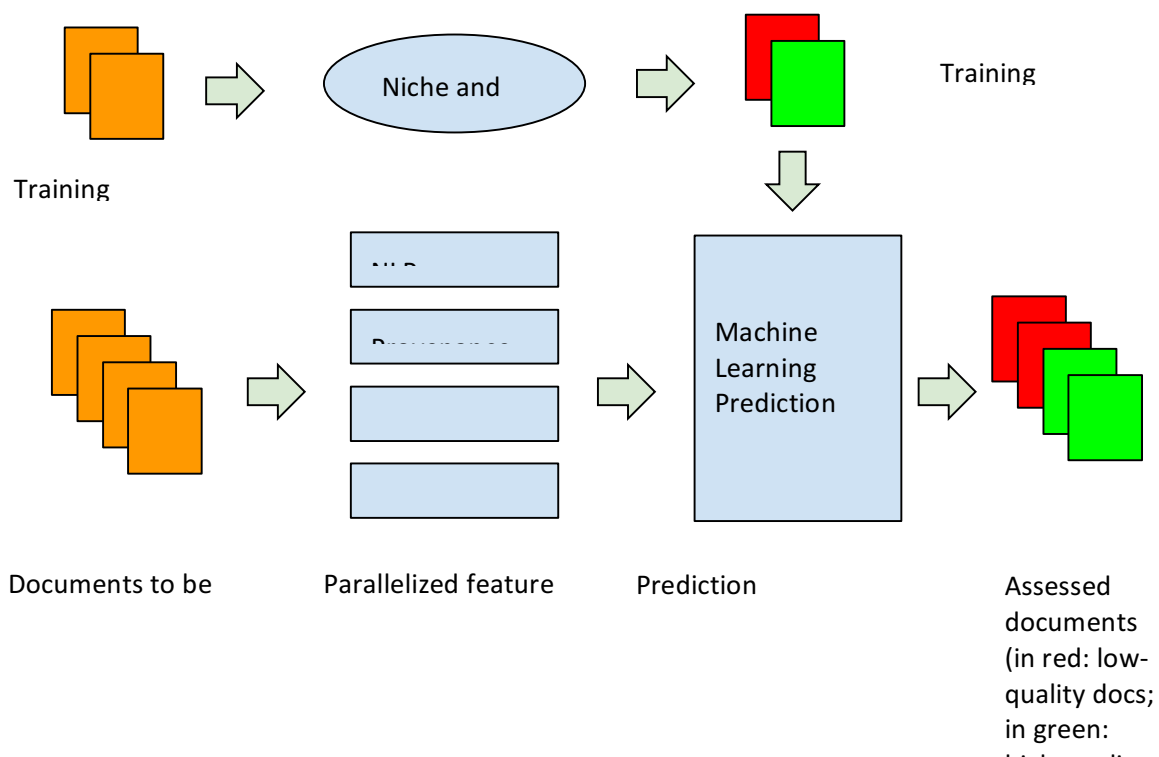


Figure 2: Automatic assessment setup.

Convolutional neural networks [4] will be evaluated as a possible solution for extending the set of documents assessed. Important will be also an analysis of the relation among the quality dimensions considered. So far, we have considered the diverse quality dimensions as independent targets to be predicted. However, it could be the case that some, lower order qualities provide the preconditions for the values of qualities of a higher order. For example, high neutrality and precision could be the necessary preconditions for high accuracy. This kind of dependencies would be favorable for improving the estimation process.

References:

- [1] D. Ceolin, J. Noordegraaf, L. Aroyo, Capturing the Ineffable: Collecting, Analysing and Automating Web Document Quality Assessments. In Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016), pages: 83-97. Springer. 2016.
- [2] V. De Boer, M. Hildebrand, L. Aroyo, P. De Leenheer, C. Dijkshoorn, B. Tesfa, G. Schreiber "Nichesourcing: harnessing the power of crowds of experts". In: *International Conference on Knowledge Engineering and Knowledge Management*. pp. 16-20., 2012. Springer.
- [3] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [4] Y. LeCun. "LeNet-5, convolutional neural networks" . Retrieved 15 February 2017.

3. Building the ARTECHNE database: New directions in Digital Art History

Marieke Hendriksen, ARTECHNE project/Department of Art History, Utrecht University
Martijn van der Klis, Digital Humanities Lab, Utrecht University

The ARTECHNE project at Utrecht University / University of Amsterdam studies how technique was transmitted among artists, artisans, and scholars between 1500 and 1950. As part of the project, the researchers are currently working with Utrecht University Digital Humanities Lab (DH Lab) to develop an online database containing searchable full-text early modern recipes, artist handbooks, and technical instructions, linked to other relevant information such as records of objects, works of art, conservation research and reconstructions. (<http://artechne.hum.uu.nl>) We aim for both quantity and quality: the more enriched texts we add, the more complex the questions we can answer using the search and visualization functions in the database.

For example, a set of questions like 'how did the use of cochineal as a pigment in oil paints change in Europe between 1500 and 1950, can we discern patterns in the spread of recipes for such paints, and are certain uses specific for particular geographical regions?' can currently only be partly answered through many years of research on primary sources such as objects and texts. Given the number of relevant sources and their limited accessibility, it will be very difficult for a researcher to discover and visualize such patterns relying on traditional art historical methods. This database, containing a great number of fully searchable and annotated sources, will allow researchers in art history, conservation, and cultural heritage to ask such complex questions and answer them with a speed and accuracy that was impossible before. Moreover, tools to detect hierarchical distance / patterns of proximity or co-occurrence of particular terms will be integrated, which can give us insight in the changing meanings of concepts.

To reach these goals, in collaboration with the DH Lab, we create the ARTECHNE database. We use Drupal (<https://www.drupal.org/>) to manage the database contents. The database is indexed using Apache Solr (<http://lucene.apache.org/solr/>), allowing researchers to use faceted search to find

relevant results in the manuscripts. The data is geotagged and contains dating information, allowing to also show search results in a GIS with an extra time dimension. Moreover, the database allows to export data from the application to .csv-format, has stable URIs and links to the Getty Vocabularies (ULAN for artist names, AAT for glossary terms and CONA for artefacts). The database thus adheres to the 5-star open data plan.

By integrating various technologies and recently developed methods in digital humanities, such as OCR, GIS, semantic annotation, crowd sourcing, and Linked Open Data in this database, we hope to firmly establish the use of enriched textual primary sources in digital art historical research, which traditionally relies heavily on images. The two authors of the paper – a historian and a scientific programmer – will present the first results of the database project. We will also reflect on the question how much digital literacy on the part of historians and how much historical literacy on the part of scientific programmers is required to successfully set up research projects relying on new technologies.

4. From Tools to “Recipes”: Building a Media Suite within the Dutch Digital Humanities Infrastructure CLARIAH

Carlos Martinez-Ortiz, Roeland Ordelman, Marijn Koolen, Julia Noordegraaf, Liliana Melgar, Lora Aroyo, Jaap Blom, Victor de Boer, Willem Melder, Jasmijn Van Gorp, Eva Baaren, Kaspar Beelen, Norah Karrouche, Oana Inel, Rosita Kiewik, Themis Karavellas and Thomas Poell

Introduction

Scholars require access to multiple, large, multimedia collections of digital resources, as well as to use a wide range of information processing tools to access and work with those collections. These requirements raise the need for developing a synchronized national and cross-national infrastructure.

Common Lab Research Infrastructure for the Arts and Humanities (CLARIAH)⁶¹ is a distributed research infrastructure for the Humanities, included on the National Roadmap for Large-Scale Research Facilities (2015-2018) drawn up by the Netherlands Organisation for Scientific Research (NWO). CLARIAH designs, implements and exploits the Dutch part of the European CLARIN and DARIAH infrastructures.

There are different research domains within CLARIAH: linguistics, socio-economic history, and media studies. Each work package within the CLARIAH project places at the centre of development both the technical requirements of each media type (text, structured data, audio-visual media), as well as the specific research needs of their user communities.

The CLARIAH Media Studies work package focuses on creating a research environment, the Media Suite (CLARIAH MS)⁶², as part of the CLARIAH infrastructure aiming to serve the needs of media scholars by providing access to audio-visual collections and their contextual data. This paper describes the approach taken to build CLARIAH MS.

Background

CLARIAH MS incorporates a series of Digital Humanities (DH) tools and aims to make them sustainable. Prototypes are currently hosted on a new infrastructure at the The Netherlands Institute

⁶¹ <http://www.clariah.nl/>

⁶² <http://mediasuite.clariah.nl/>

for Sound and Vision (NISV) data centre. These prototypes are: AVResearcherXL, TROVe, CoMeRDa, Oral History Today (OHT) and DIVE+. Furthermore, CLARIAH MS aims to support audio-visual archives in opening up collections in a more standardized way. Once these objectives have been accomplished, scholars will be able to search and analyse these collections via a central workspace, thus, enabling *data intensive research* in the humanities.


AVResearcherXL is an exploratory tool which enables simultaneous queries and analytic visualizations of the collections' metadata (Van Gorp et al (2015)). TROVe was developed to ease the combined access and visualization of archival collections and online social media. CoMeRDa is a web based aggregated search system for visualizing search results (Bron et al(2013)). OHT is a prototype for search and enrichment (through Automatic Speech Recognition technology) of distributed Oral History collections in The Netherlands (Ordelman and de Jong(2011)). Finally, DIVE+ is a linked-data digital cultural heritage collection browser which provides access to heritage objects from heterogeneous collections, using historical events and narratives as context for searching, browsing and presenting the objects (de Boer et al.(2015)).

These five tools support scholars in the “exploration” and “contextualization” phases of their research, a framework proposed in (Bron et al.(2015)). The original tools could not interoperate and did not operate on the same data, which limits their potential. Recreating them in a single configurable environment makes it possible to reuse functionalities across data sets and to reuse data across functionalities.

CLARIAH Media Suite


The DH community includes scholars with a wide diversity of research interests and goals; every research group in DH is working with different types of data and their research objectives have specific requirements which cannot be easily facilitated by tools using a single, generic approach. Simultaneously, there are similarities in the methods used by different scholars (de Jong et al.(2011)) that can be used for generalised tool development. There are commonalities in research questions and methods among media scholars, which we grouped into *Media aesthetics*, *Social history of media*, *Aesthetic historiography*, *Social and cultural history*, *Media representations or coverage*, *Transmedia analysis*, and *Memory studies* (Melgar et al., 2017).

CLARIAH Media Suite
Data sources
APIs
Components
Recipes
Help & Feedback
Login




What is the Media Suite?

The Media Suite is a research environment of the Dutch infrastructure for digital humanities and social sciences (CLARIAH) which aims to serve the needs of **media scholars** (and other scholars who use audiovisual media) by providing access to audiovisual collections and their contextual data. Here you can enter the Media Suite services. To read more about the **CLARIAH Media Studies focus**, go to the CLARIAH website, or to our **Media Suite blog and documentation pages**.




Data Sources

Collections are registered in a common inventory which describes their metadata and are available in Elasticsearch and RDF format.




APIs

The foundation of the Media Suite is built on our APIs. They facilitate the interaction with data from various collections.



Components

Each component performs a single, well-defined task. Modules can interoperate to construct more sophisticated functionality.



Recipes

Recipes are built based on the knowledge and experience of the scholarly community. They integrate components into tools for media studies research.

Figure 1. CLARIAH MS consists of functionalities, APIs and recipes, version 1, April 2017

A generic infrastructure is required to cater for the general needs of every user group. The infrastructure needs to incorporate flexible functionality capable of addressing very specialized research questions. Media scholars expressed their desire to use the collections and tools which were previously “locked” together in the individual prototypes. CLARIAH MS has been designed in a modular way (Figure 1); each module performs a single, well-defined task. Modules can interoperate to construct more sophisticated functionality.

Metaphorically speaking: whereas previously users had access to predefined ‘meals’ - tools which could perform cross-collection search and visualize the results in the form of timelines, word clouds, snippets and/or thumbnails - we now provide users with single ingredients (individual functionalities such as searching), and ready-made recipes (combinations of several functionalities). Some ingredients may be used in different recipes, existing recipes may be complemented by adding extra ingredients.

Media Suite Architecture

CLARIAH MS consists of four layers of functionality, explained below (Figure 2):

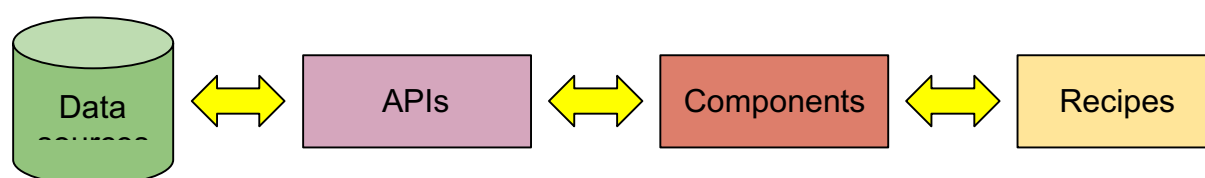


Figure 2 - Architectural design of CLARIAH MS.

Data Sources contain the collections (e.g., television broadcasts from NISV, EYE Jean Desmet collection, DANS Oral History collection). All collections are registered in a common inventory (CKAN⁶³) which describes their metadata. Collections are available in Elasticsearch (full text search) and RDF format (semantic search).

APIs facilitate the interaction with data from various collections:

- Collections API - high-level collection information (metadata: data format, size, etc.)
- Search API - searching for collection items.
- Annotation API - annotating existing data using W3C Web Annotation standard (mainly for manual annotations)(Melgar et al.,(2017)).
- Data Enrichment API - collection enrichment through automatic mechanisms (e.g. name entity recognition) or by human interaction (e.g. crowdsourcing).

The APIs design allows the integration of new data of different formats and data models.

Components in CLARIAH MS are software units which perform a single functionality: each component takes data as input and produces a meaningful output using standard formats, to be connected with other components (e.g., word cloud, timeline visualizations, topic identification in newspapers, searching content in collections).

Recipes close the circle by integrating components to recreate the functionalities of the original tools. We focus on providing the complex functionality of the original tools in the form of four

⁶³ <http://mediasuite.clariah.nl/datasources>

'recipes'. Following with the metaphor above, the concept of ingredients (components) allows researchers to prepare their own personal recipes (functionalities).

Conclusion

In this paper we have explained the structure of the CLARIAH MS and how previously developed DH tools are being integrated in a sustainable infrastructure that allows flexible use of data collections and functionalities fitting the research needs of scholars. We have also sketched our strategy to enable the integration of alternative functionalities and data collections using a modular approach (ingredients and recipes). Future work includes user evaluation of the first version of the Media Suite (launched in April, 2017), and co-development involving six CLARIAH research pilot projects⁶⁴.

References

- [Bron et al. (2013)]** Marc Bron, Jasmijn Van Gorp, Frank F. Nack, Maarten de Rijke, Lotte B. Baltussen. *Aggregated search interfaces in multi-session tasks*. SIGIR 2013: 36th international ACM SIGIR conference on research and development in information retrieval. Dublin: ACM (2013)
- [Bron et al.(2015)]** Marc Bron, Jasmijn Van Gorp, and Maarten Rijke. Media studies research in the data-driven age: How research questions evolve. *Journal of the Association for Information Science and Technology* (2015), <https://doi.org/10.1002/asi.23458>.
- [de Jong et al.(2011)]** Franciska de Jong, Roeland Ordelman, and Stef Scagliola. *Audio-visual collections and the user needs of scholars in the humanities: a case for co-development*. In *Proceedings of the 2nd Conference on Supporting Digital Humanities (SDH 2011)*, Copenhagen, Denmark, 2011. Centre for Language Technology, Copenhagen.
- [Melgar et al.(2017)]** Liliana Melgar Estrada, Marijn Koolen, Hugo Huurdeman, and Jaap Blom. *A process model of time-based media annotation in a scholarly context*. In *ACM Conference on Human Information Interaction and Retrieval (CHIIR)*, Oslo, 2017.
- [Ordelman and de Jong(2011)]** Roeland Ordelman and Franciska de Jong. *Distributed access to oral history collections: Fitting access technology to the needs of collection owners and researchers*. In *Digital Humanities 2011: Conference Abstracts*, pages 347–349, Stanford, 2011. Stanford University Library. URL <http://purl.utwente.nl/publications/78347>. ISBN=978-0-911221-47-3.
- [de Boer et al.(2015)]** Victor de Boer, Johan Oomen, Oana Inel, Lora Aroyo, Elco van Staveren, Werner Helmich, Dennis de Beurs: *DIVE into the event-based browsing of linked historical media*. *J. Web Sem.* 35: 152-158 (2015)
- [Van Gorp et al (2015)]** Jasmijn Van Gorp, Sonja de Leeuw, Justin van Wees, Bouke Huurnink. *Digital Media Archaeology - Digging into the Digital Tool AVResearcherXL*. *VIEW. Journal of European Television History and Culture/E-journal*, 4 (7): 38-53 (2015)

⁶⁴ <http://www.clariah.nl/projecten/research-pilots>

1. Digitally mediated emotions: representations and reinforcements

Anca Țenea - Doctoral School "Space, Image, Text, Territory" – CESI

In 2014, a research conducted by Facebook analyzed 689.003 users' news feeds, by delivering more positive or negative content to each user (Kramer et al. 8788). The conclusions stated the contagious factor of emotions in digital space, by showing that people who were provided, say, more positive content, tended to express more positivity in return, by distributing themselves positive messages. The assumptions that machines stimulate emotions in order to enhance them was thus reconfirmed. Yet, there are still many questions emerging from the current state of digital media.

The research project I am proposing analyses the connection between digital media and human emotions, as users participate in a digital spectacle in which they get their active part through their emotions. The manifestations of their emotions are read and interpreted by the platforms' algorithms in order to respond by providing a spectacle according to the user's desires, beliefs, and actions. I propose a theoretical approach to the way human emotions, feelings, and affects are mirrored and enacted in digital space. In light of the new theories regarding the importance of sentiment mining and affective computing in shaping human knowledge, affects and behavior, I argue the necessity of analysing how emotions are stimulated in networked publics (Boyd) in order to enhance the participation to the digital spectacle.

The digital spectacle addressed in this research refers to the collection of information that returns to the user in the form of personalized content, in response to their online activity, and to the information they display via the Internet. Transforming emotions into data constitutes a pivotal mechanism for digital technology, where the user is not only the spectator but a fully engaged participant. Interestingly, I argue, this has the potential to reveal the mechanism through which the user relates to the Lacanian Other in an online process of virtual identity formation. The elements that trigger a Freudian drive and encourage the interaction with the Lacanian Other, as well as the characteristics of digital platforms meant to provide fantasmatic and identitary projections will also be examined.

My research focuses on a series of platforms and softwares that ensure such interactions. For instance, I am interested in how the Persado cognitive mechanisms are set to detect emotions through sentiment mining, for the benefit of commercial advertising. Persado identified 16 emotions as triggers for user action, and the case studies posted on their website show that their strategy increasingly improved marketing performance for different campaigns. A question occurs: how are these commercial messages becoming triggers for user reaction?

I am also enquiring into Facebook's reaction buttons and "On this day" feature effects on sentiment, affect and emotion reinforcement. Emotions are deeply connected and influenced by social norms, which dictate how one should feel, and by behavioral codes, which influences people on expressing emotions (Benski and Fisher 3). I argue that Facebook, by means of its architecture, is built to enhance feelings and affects, and expose a

user to both social norms and behavioral codes, inducted by other people's posts, magazine and brands posts and ads. The reactions buttons stimulate their emotions to different situations: like, love, angry, sad and wow are very similar to the ones defined by Paul Ekman in 1972, in "Emotion in the Human Face" (surprise, fear, sadness, happiness, disgust and rage), as key and prevalent emotions in human behavior.

The reactions symbols also partly correspond to emotions postulated by Jacques Lacan and Melanie Klein, as fundamental resorts of identity construction, such as fear, pleasure, anxiety, fury, joy etc. The digital preponderant emotions can be linked to the good and bad objects (Melanie Klein), through the manifestation of the things we observe and interact with from the digital screen. I would argue that even though the speed of online emotional reactivity is higher than in real life interactions, the operative mechanism is basically the same.

This opens another discussion, on the visual triggers in social media, through psychoanalytical theories. Since the digital screen is composed of objects of desires, or Lacanian "objets petit a", the digital interfaces can be investigated through visual pleasure and identification. Technology is often compared with fetishism, which involves syncing different symbols with objects or pleasure. This idea - states Andre Nusselder (19) - is supported by the fact the technologies transcend the limits of regular life, offering pleasure, opening endless possibilities. He describes this aspect as a hallucinatory imagination of reality because digital technologies synchronize humans with the pleasure principle, postulated by Freud and reinforced by further psychoanalytical theories as the motor of human pleasure. The spectator is therefore no longer passive, as they continuously interact with the screen, influencing the content. Images and symbols online represent the objects of desire, which are part of the Imaginary that simulates and stimulates, creating the digital fantasy.

Platforms aim to offer users a spectacle compatible to their conceptual apparatus, reinforcing familiar mythologies and beliefs, as well as their registered common desires. In the light of the questions on why do people react to certain content, be it on social media, newsletters or other particular advertising messages, I find it legitimate to ask which are the triggers that make a user take an action.

Works Cited

Boyd, Danah (2010). "Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications." *Networked Self: Identity, Community, and Culture on Social Network Sites* (ed. Zizi Papacharissi), 2010, pp. 39-58, www.danah.org/papers/2010/SNSasNetworkedPublics.pdf.

Fisher, Eran; Benksi, Tova, *Internet and Emotions*: Routledge, New York, 2014

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks." *Social Sciences - Psychological and Cognitive Sciences*, vol. 111, no. 24, pp. 8788-8790, www.pnas.org.

Nusselder, André, *Interface Fantasy: A Lacanian Cyborg Ontology*, Cambridge: The MIT Press Cambridge, 2009.