TwinTalks at DHN 2018 – Understanding Collaboration in Digital Humanities

Steven Krauwer[1,2] and Darja Fišer[1,3]

[1] CLARIN ERIC, Utrecht, Netherlands
[2] Utrecht University, Netherlands
[3] University of Ljubljana, Slovenia
s.krauwer@uu.nl
darja.fiser@ff.uni-lj.si

Abstract. In this preface to the workshop proceedings we describe briefly the rationale and the format of the workshop.

Keywords: Digital humanities. Collaboration. Training.

Rationale and format.

Humanities research can only take full advantage of new technological developments if content and digital expertise work hand in hand, very similar to the hard sciences where research is done in teams, where all the members work on a specific problem by bringing in his/her expertise and skills, be it content-related or technical. It is becoming clear that co-design, co-development and co-creation should be the norm rather than the exception in the humanities as well, but very little is known about how this collaboration works in practice and how better training and education of both humanities scholars and digital experts could facilitate such collaboration. This is what this workshop addresses, based on real life collaboration examples. In particular, we invited researchers, professionals, educators, and RI operators with a special interest in creating the conditions where humanities scholars and technical experts can fruitfully collaborate in answering humanities research questions.

The main objective of the workshop was to get a better understanding of the dynamics on the digital humanities work floor where humanities scholars and digital experts meet and work in tandem to solve humanities research questions. The best way to do this seems to be to give both parties the opportunity to present their achievements and share their collaboration experiences with the audience. The insights gained should help those involved in the education of humanities scholars, professionals and technical experts alike to develop better training programmes.

This workshop is special in that all papers in this workshop were submitted and (as far as possible) presented in tandems of a humanities researcher and a digital expert. They

reported on the research carried out together, both from their individual perspective (either humanities research or technical), as well as on their collaboration experience.

The programme started with an invited talk by Mikko Tolonen, which was then followed by two long and six short presentations. The talks had to contain the following three components: presentation of the humanities research problem and its solution, presentation of the technical aspects of the research done, and a report on the collaboration experience itself, including the obstacles encountered and recommendations on how better training and education could help to make the collaboration more fruitful. The programme ended with a round table discussion with all the participants in order to summarize the lessons learned from the presentations.

In order to reach a broad audience all humanities research topics in a very broad sense were welcome, where we explicitly included social sciences as well as cultural heritage studies. The research could be completed or ongoing, as long as the presentation explicitly addressed the way the humanities researcher and the digital expert have collaborated or still collaborate. For this latter point we asked authors to address issues such as (but not limited to):

- What was easy and what was difficult – and why?
- How did the researcher and technician change each other's way of looking at things?
- Did they, for instance, make each other aware of blind spots they had?
- Did the combination of thinking from a DH research question and thinking from a technical solution lead to new insights?
- How could better training or education of scholars and digital experts make collaboration easier, more effective and more efficient?

As the programme shows, a wide variety of topics was covered in the workshop, ranging from using computer vision for classification of historical newspaper images to making cultural content in non-standard language available for cross-disciplinary research. In total, 28 authors from 7different countries contributed to the workshop, extending well beyond the central focus of the DH Nordic conference, which only confirms that this discussion is both much needed and appreciated in our community.

Steven Krauwer and Darja Fišer
Utrecht and Ljubljana, 18 February 2019

Workshop programme

Workshop TwinTalks: Understanding collaboration in DH

Programme

Steven Krauwer and Darja Fišer: *Welcome and introduction*

Mikko Tolonen (invited talk):

Why humanities research questions should come first? Reflections on different kinds of collaboration in digital history

Martijn Kleppe, Thomas Smits and Willem Jan Faber:

Three perspectives on a collaborative attempt to use computer vision techniques to automatically classify historical newspaper images

Konstantin Freybe, Florian Rämisch and Tracy Hoffmann:

With small steps to the big picture - A method and tool negotiation workflow

Alptug Güney, Cristina Vertan and Walther von Hahn:

Combining hermeneutic and computer based methods for investigating reliability of historical texts

Börge Kiss, Daniel Kölligan, Francisco Mondaca, Claes Neuefeind, Uta Reinöhl and Patrick Sahle:

It Takes a Village: Co-developing VedaWeb, a Digital Research Platform for Old Indo-Aryan Texts

Vanessa Hannesschläger and Peter Andorfer:

I want it all, I want it now: Literature researcher meets programmer

Eetu Mäkelä, Mikko Tolonen, Jani Marjanen, Antti Kanner, Ville Vaara and Leo Lahti:

Exploring the Material Development of Newspapers

Maria Papadopoulou and Christophe Roche:

Twinning Classics and A.I.: Building the new generation of ontology-based lexicographical tools and resources for Humanists on the Semantic Web

Amelie Dorn, Yalemisew Abgaz and Eveline Wandl-Vogt:

Opening up cultural content in non-standard language data through cross-disciplinary collaboration: insights on methods, processes and learnings on the example of exploreAT!

Discussion (All)

Steven Krauwer and Darja Fišer: *Wrapping up and closing*

Programme committee

This workshop was a joint initiative by CLARIN ERIC (www.clarin.eu), DARIAH-EU (www.dariah.eu) and the PARTHENOS project (www.parthenos-project.eu).

Chairs and main organisers:
Steven Krauwer (CLARIN ERIC / Utrecht University; steven@clarin.eu)
Darja Fišer (CLARIN ERIC / University of Ljubljana; darja.fiser@ff.uni-lj.si)

PC members:
Franciska de Jong (CLARIN ERIC / Utrecht University)
Bente Maegaard (CLARIN ERIC / University of Copenhagen)
Jennifer Edmond (Trinity College Dublin / PARTHENOS / DARIAH-EU)
Ulrike Wuttke (University of Applied Sciences Potsdam / PARTHENOS)
Frank Uiterwaal (NIOD – KNAW / PARTHENOS)
Eleni Gouli (Academy of Athens / PARTHENOS)
Koenraad De Smedt (University of Bergen, CLARINO)

Three perspectives on a collaborative attempt to use computer vision techniques to automatically classify historical newspaper images

Martijn Kleppe[1][0000-0001-7697-5726] Thomas Smits[2][0000-0001-8579-824X] Willem Jan Faber[3]

[1+3] National Library of the Netherlands, The Hague, the Netherlands
[2] Utrecht University, Drift 6, Utrecht, The Netherlands
Martijn.Kleppe@KB.nl

Abstract. In the last couple of years, scholars in the Humanities have started to explore the possibilities of the large-scale analysis of images. This development can be linked to the increasing availability of large visual datasets, the increase in computing power, and the development of new techniques, such as convolutional neural networks. However, there are no one-size-fits all researchers that are able to gather the right data, apply the new techniques, and analyze the results in meaningful ways. In this paper we present the collaboration of a Humanities researcher, a Research Software Engineer and Digital Scholarship Advisor to explore how new computer vision techniques can be used to automatically classify images extracted from a large collection of digitized historical newspapers. We will present the outcomes of our research and share the lessons we learned from our collaboration. First we will discuss the experiences of the Humanities researcher. Second we will discuss the lessons we learned from a technical perspective. Third, we will elaborate on the institutional perspective of the National Library of the Netherlands (KB) as a data provider but also as full partner of the research project. We will end with a reflection on the broader strategic role of heritage institutes as research partners to stimulate, collaborate and to preserve results of research projects in a sustainable manner.

Keywords: Computer Vision, Distant viewing, Digitized newspapers

1 Introduction

Although the Digital Humanities have traditionally focussed on the large-scale analysis of texts (Nicholson, 2013), recent years have seen an upsurge in research that focuses on images. This move to the visual can be explained by the increasing availability of visual datasets (Russakovsky et al., 2015) and the techniques necessary to analyse them. Examples of this kind of research include the work of Seguin (Seguin et al., 2017) who focuses on automatic visual pattern detection across iconographic collections and the work of King and Leonard (2017) on colometrics, facial detection and neural network-based visual similarity. The International Digital Humanities conferences also displayed a growing interest for non-textual sources (Weingart, 2016), reflected in the workshops on computer vision organised by the Special Interest Group Audiovisual Material in Digital Humanities in 2017 (Kleppe et al., 2017) and 2018 (Tilton et al., 2018).

In the Netherlands we see a similar tendency. First, datasets of digitised visual sources are becoming more available. The National Library of the Netherlands (KB) offers access to a large collection of digitised newspapers on their portal www.delpher.nl, allowing full-text searches through all data and drill down the results by applying filters such as period, region or type of article. Furthermore, researchers can get access to all digital sources through the library's Dataservices and APIs and experimental datasets at the KB Lab, such as the KBK-1M Dataset (Kleppe et al., 2016). To stimulate the use of these datasets, understand the needs of researchers, and improve the library's services, the KB has set up the researcher-in-residence program (Wilms, 2017; Boekestein, 2017). This allows researchers to work part time at the Research Department of the KB for six months, together with one of KB's Research Software Engineers. During their project they are also assisted by a Digital Scholarship Advisor and several metadata and collection specialists.

In 2017, two researcher-in-residence projects were carried out to explore the possibilities of applying new computer vision techniques to analyse digitised historical newspapers. Melvin Wevers explored visual similarity search on newspaper advertisements (Wevers and Lonij, 2017). In this paper, we will focus on the second project by Thomas Smits, on classifying newspaper images. We will first describe the Humanities research question, followed by our technical approach and the project's results. The final part of the paper is a reflection on the collaboration between the Humanities Researcher and the Research Software Engineer. We will also reflect on the role of the KB as a data provider, but also full research partner.

2 Humanities research question: Fin de siècle visual news culture

The visual representation of news events is generally connected to the technological progress of photography (Gervais and Morel, 2017). The so-called half-tone revolution of the early 1880s, enabling the massive reproduction of photographs in print media, is seen as forming the basis for our current visual news culture. Several historians of nineteenth-century media have challenged this narrative (Gitelman and Pingree, 2003). Hill and Schwartz (2015) propose a contingent history of 'news pictures' as a separate 'class of images', which not solely focuses on photographic technology, but on the discourse surrounding them (p. 3). In relation to this recent theoretical development, several studies have demonstrated that photography was not the first medium used to visually represent the news. From the early 1840s, illustrated newspapers disseminated news pictures on a massive scale and developed a discourse of objectivity, based on eyewitness accounts, which would be adapted and used for photographs later in the century (Barnhurst and Nerone, 2000; Gervais, 2010; Park, 1999).

Although the visual representation of the news did not start with photography, the pre-eminence of this medium is clear in the twentieth century (Gervais and Morel, 2017; Kester and Kleppe, 2015). It follows that the turning point between the use of illustrations and photographs as the preferred medium to represent the news is a critical moment in the history of modern visual news culture. Most commonly, researchers have presented this point as a watershed, located at the publication of the first photograph of a news event in a newspaper (Kester and Kleppe, 2015). However, case studies from a media archaeological perspective, suggest a relatively long transitional

period in which illustrations and photographs coexisted and competed as authentic, objective visual representations of the news (Keller, 2013; Steinsieck, 2006). It remains unclear when photography exactly achieved its pre-eminence and why this happened.

The earlier reliance on case studies to describe the transitional phase is understandable, as, in pre-digital times, a 'distant reading' (Moretti, 2015) of the large number of images published in newspapers was all but impossible. Using several computer vision techniques, our project aspired to shed more light on this important debate by analysing pictures of the news in Dutch newspapers from a distance ('distant viewing') and on a large scale. Our main research questions were: When did Dutch newspapers start to use illustrations? And when did they switch to using photographs as the primary visual medium? More generally, we hoped to explore how these techniques could be used to analyse large collections of visual historical material.

3 Technical approach: convolutional neural networks

As most DH research, our project faced two main challenges: data collection and data analysis. Within Delpher, users can select facets to drill down to specific results. Upon selecting 'Illustration with caption' they will only get articles that contain an image. However, the results will not only contain photographs, but also cartoons, drawings, weather reports and even graphic displays of chess problems. Since this would not suffice to answer our main research question, we had to find new ways to classify the images found in newspapers.

Concerning data collection part, our project could build on the PhoCon project of Elliott & Kleppe (Kleppe et al., 2016), which created a database containing images extracted from Delpher's newspapers. However, the result of this project, the KBK-1M(illion) database only contained images from the period 1923-1930. Furthermore, we found that not all images in the period of our research (1860-1923) were correctly classified as 'captioned illustration' by the OCR company. Therefore, new code was needed to harvest all the images from digitised newspapers. We found that in the XML files (ALTO) the code-line 'imageblock' denotes images. Around 1900, Dutch newspapers contained many small images, like the often-recurring illustrations used at the beginning of a specific section, or small images that accompanied advertisements in newspapers. Because we were mainly interested in images of the news, we decided to only include images that could be related to newspaper articles (via the XML file), exclude images of advertisements, and discard all the images with a file size smaller than 30KB. We ended up with 313K images for the period 1923-1930.

We classified these images using a three-step pipeline. First of all, we used Adam Geitgey's facial recognition API, built using the Dlib's facial recognition library, to recognize faces on the images (Geitgey, 2017). In the second step, using the 'Tensorflow for poets' method, we applied an Inception-V3 convolutional neural network to recognize nine different categories (buildings, cartoons, chess, crowds, logos, maps, schematics, sheet music, and, weather reports).[1] Although the creators of this method recognize that it will be outperformed by a full training run, it is surprisingly effective (see below for performance) and does not require GPU hardware. We used training sets of around forty images for every category. For the final classification step, we

[1] https://codelabs.developers.google.com/codelabs/tensorflow-for-poets/#0

asked Leonardo Impett, a digital art historian at the Bibliotheca Hertziana, to build a convolutional neural network that could recognize if images were either drawings or photographs. His network focuses on the lower-layers of the network and a support vector machine (SVM) divides the images into photographs and illustrations.

The four-step classification pipeline resulted in the CHRONIC (Classified Historical Newspaper Images) database, which contains metadata for all the 313K newspaper images we extracted (Smits and Faber, 2018a). Based on this database, we created CHRONReader: a tool which allows users to search for images containing faces, one of the nine categories and being either illustrations or photographs (Smits and Faber, 2018b).

4 Results

Using computer vision we were able to analyse the images of Dutch newspapers on a large scale, or view them from a distance, and, as a result provide an answer to the main research questions: When did Dutch newspapers start to use illustrations? And when did they switch to using photographs as the primary visual medium? Fig. 1 depicts the publication of illustrations and photographs in Dutch newspapers between 1860 and 1930. The number of images in Dutch newspapers, both illustrations and photographs, increased noticeably in the early 1900s and peaked at the start of the 1920s. The number of photographs overtook the number of illustrations for the first time in 1927. This completed a development from nineteenth-century publications filled with letters, to pages filled with both images and text: the form of the newspaper we still know today.

On the one hand, the application of CNNs thus confirms the conclusions of earlier work, mentioned above, based on case studies. At the same time, vast digitized archives and new techniques like CNNs contribute to the construction of a exciting new overview of visual (news) culture, which allows for the analysis of trends and changes over an extended period of time. As Fig. 1 shows, the visual representation of the news took off in the earlier 1920s. Although earlier research noted and analysed the introduction of so-called ‘photo-pages’ in the 1920s using a limited set of sources (Broersma, 2014; Kester and Kleppe, 2015) the birds-eye view of the use of images in the entire Dutch press provides us with a new perspective on the magnitude of this watershed moment.

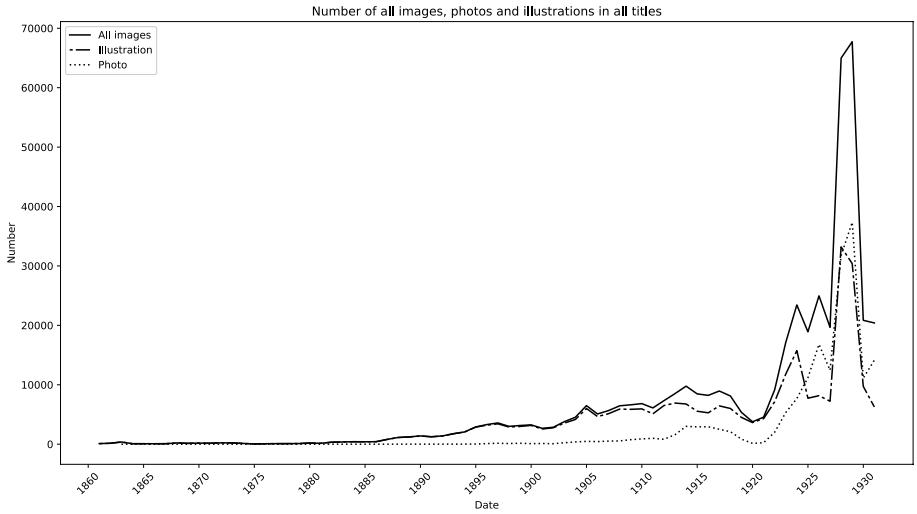

Fig. 1. Number of all images, photo's and illustrations in all digitized newspapers, 1860-1930

Next to the ability to view large collections of images from a distance, new computer vision techniques also provide direct access to visual content without having to refer to textual descriptions. In this sense the technique can be compared to OCR-technology, in that it provides users of digital archives with bottom-up access to sources (Nicholson, 2013).

For the KB, this project offered several results. First we gained more knowledge about the user needs of researchers who want to study the visual aspects of digital sources. Second, we got to know our data and metadata better. For example, resulting from the set-up of the metadata and the way this is created and stored at the KB, the creation of a dataset containing images and their captions turned out to be more complicated than expected. Third, due to the collaborative nature of the researcher-in-residence program, our research software engineer gained more knowledge about computer vision. Since libraries have been focused on texts for centuries, we nowadays mainly focus on Natural Language Processing techniques to analyze digital material. However, as we have learned from this and the PhoCon project, digital datasets also contain millions of images. Since libraries continuously want to improve access to their digital collections, they should focus on both textual and visual material. However, as we have learned from this project, this is far from an easy task. The assistance by Leonardo Impett to build a convolutional neural network to divide the images into photographs and illustration was e.g. fundamental for the end result of the project. Fourth, since the KB now has the knowledge on applying computer vision, we are taking steps to apply it on a large scale. On Delpher.nl users can select 'Illustrations with captions' but as we have described before, they then retrieve all sorts of images. The results of the CHRONIC project allows the KB to classify all images in historical newspapers to eventually implement an advanced selection option within Delpher to allow users also to select photographs, cartoons or even chess problems. However, scaling up the results

of this research project to the full collection will present several challenges in terms of computing power and infrastructure

5 Reviewing collaboration

For this project, we set up a team consisting of a Humanities researcher (Thomas Smits), a research software engineer (Willem Jan Faber) and digital scholarship advisor (Martijn Kleppe). The team met on a weekly basis to discuss the projects' progress, while the individual team members also regularly had bilateral meetings or were helped by KB's in house metadata and collection specialists. Since the Humanities researcher was researcher-in-residence, he was seconded for six months to the KB and was present in the KB for two days a week, which was very stimulating for the projects progress. He could easily get access to KB's in house experts who normally can only be contacted through KB's front office. In this way, he was able to get more easy access to (meta)data and more specialised knowledge about the data structure. Furthermore, the collaboration with the research software engineer allowed him to explore not only the data but also new techniques. Trained as a traditional historian, Smits was not used to working with innovative, and highly complex, digital methods of analysis, such as neural networks. Due to the intensive nature of the collaboration within the researcher-in-residence program, he eventually was able to understand the techniques applied and extrapolate them to the results of the project.

For the KB, this is a pivotal project showing the added value of close collaboration with a researcher. Although the KB participates in many research projects, its main role is acting as data provider, allowing researchers to use the large datasets of the KB. However, the library can do more to take full advantage of the knowledge created in these projects and implement the results of the research to its collections. Given the collaborative nature of the researcher-in-residence program, both aspects are covered. Since Smits is a domain expert in the field of historical visual culture, he helped the KB to understand their data better and together with the research software engineer he created a training set to build the algorithm that classified the images. If the KB manages to apply this algorithm to all images in the KB dataset and implement the filter option in Delpher, the results of this collaboration will be beneficial to all visitors of www.delpher.nl.

6 Conclusion

The project was a success for all parties involved. The Humanities researcher was able to answer his main research question and presented the results at several conferences (Smits, 2017; Smits en Wevers, 2018a, 2018b) and published an article in *Digital Scholarship in the Humanities* (Wevers and Smits, 2019). Furthermore, the Humanities researchers and the research software engineer created a dataset, tool and code that are all freely available through KB's Lab. The research software engineer gained a lot of knowledge about the possibilities of computer vision techniques to further open up the libraries digital collection. Finally, the digital scholarship advisor is currently exploring the possibilities to implement the results of the project within Delpher so that it can benefit a large audience (Delpher.nl has two million visits per year).

This last conclusion is an example of the potential of applying research results to library services in order to open up digital collections to a wider audience. Earlier, Peter Leonard (2016) made a plea for this for this when he stated he wanted to 'put TDM in the mainstream.' Alex Humphreys (2018) made a similar plea for 'Applied Digital Humanities' and (Kleppe, 2018) also referred to the potential of 'Libraries as incubators for DH Research Results'. It demonstrates the crucial role institutes, such as libraries, can play within research projects. When these institutes go beyond the role of data provider, they are not only a full partner by bringing and gaining knowledge, but they can also act as the ideal valorisation vehicle of research projects. By taking up an active role in adopting relevant research results in their own services, they can preserve these results in a sustainable manner and bring the affordances of DH research to the wider public.

References

Barnhurst, K., Nerone, J., 2000. Civic Picturing vs. Realist Photojournalism. The Regime of Illustrated News, 1856-1901. Design Issues 16, 59–79.

Broersma, M., 2014. Vormgeving tussen woord en beeld. De visuele infrastructuur van Nederlandse dagbladen, 1900 – 2000. Tijdschrift voor Mediageschiedenis 7, 5–32.

Geitgey, A., 2017. Face_recognition: The world's simplest facial recognition api for Python and the command line: https://github.com/ageitgey/face_recognition

Gervais, T., 2010. Witness to War: The Uses of Photography in the Illustrated Press, 1855-1904. Journal of Visual Culture 9, 370–384.

Gervais, T., Morel, G., 2017. The Making of Visual News: A History of Photography in the Press. Bloomsbury Academic, London.

Gitelman, L., Pingree, G. (Eds.), 2003. New Media, 1740-1915. MIT Press, Cambridge.

Hill, J., Schwartz, V., 2015. Getting the picture: the visual culture of the news. Bloomsbury Academic, London.

Humphreys, A., 2018. The Case for Applied Digital Humanities in Scholarly Communications. Presented at the SSP Annual Meeting, Chicago.

Keller, U., 2013. The iconic turn in American political culture: speech performance for the gilded-age picture press. Word and Image 29, 1–39. https://doi.org/10.1080/02666286.2012.729794

Kester, B., Kleppe, M., 2015. Persfotografie. Acceptatie, professionalisering en innovatie, in: Bardoel, J., Wijfjes, H. (Eds.), Journalistieke Cultuur in Nederland. Amsterdam University Press, Amsterdam, pp. 53–76.

King, L., Leonard, P., 2017. Processing Pixels: Towards Visual Culture Computation. Presented at the ADHO 2017.

Kleppe, M., 2018. Keynote: Bringing Digital Humanities to the wider public: libraries as incubator for DH research results. Presented at the Language Technologies & Digital Humanities Conferences, Ljubljana, Slovenia. https://doi.org/10.5281/zenodo.2532678

Kleppe, M., Elliott, D., Faber, W.J., 2016. *K*oninklijke Bibliotheek Kranten – 1 Miljoen (KBK-1M). KB Lab: The Hague. http://lab.kb.nl/dataset/kbk-1m // https://doi.org/10.17026/dans-xar-hqvg

Kleppe, M., Lincoln, M., Wevers, M., Williams, M., Seguin, B., Smits, T., 2017. Computer Vision in Digital Humanities, in: Confernce. Presented at the DH2017, ADHO, Montreal, pp. 833–836.

Moretti, F., 2015. Distant reading. Verso, London.

Nicholson, B., 2013. The Digital Turn. Media History 19, 59–73.

Park, D., 1999. Picturing the War: Visual Genres in Civil War News. The Communication Review 3, 287–321.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211–252.

Seguin, B., di Leonardo, I., Kaplan, F., 2017. Tracking Transmission of Details in Paintings. Presented at the Digital Humanities 2017, Montreal.

Smits, T., 2017. Illustrations to Photographs: using computer vision to analyze news pictures in Dutch newspapers, 1860-1940. Presented at the Digital Humanities 2017, Montreal.

Smits, T., Faber, W.J., 2018. CHRONIC (Classified Historical Newspaper Images). KB Lab: The Hague. http://lab.kb.nl/dataset/chronic-classified-historical-newspaper-images

Smits, T., Faber, W.J., 2018. CHRONReader. KB Lab: The Hague. http://lab.kb.nl/tool/chronreader

Smits, T., Wevers, M., 2018a. Seeing History: The Visual Side of the Digital Turn. Presented at the DH2018, Mexico City.

Smits, T., Wevers, M., 2018b. Seeing History: The Visual Side of the Digital Turn. Presented at the DHBenelux 2018, Amsterdam.

Steinsieck, A., 2006. Ein imperialistischer Medienkrieg. Kriegsberichterstatter im Südafrikanischen Krieg (1899–1902), in: Daniel, U. (Ed.), Augenzeugen. Kriegsberichterstattung vom 18. zum 21. Jahrhundert. Vandenhoeck & Ruprecht, Göttingen, pp. 87–112.

Tilton, L., Arnold, T., Smits, T., Wevers, M., Williams, M., Torresani, L., Bell, J., Latsis, D., 2018. Computer Vision in DH. Presented at the DH2018, Mexico City.

Wevers, M., Lonij, J., 2017. SIAMESET. KB Lab: The Hague. http://lab.kb.nl/dataset/siameset

Wevers, M., Smits, T., 2019. The Visual Digital Turn. Using Neural Networks to Study Historical Images. Digital Scholarship in the Humanities (accepted).

With small steps to the big picture
A method and tool negotiation workflow

Konstantin Freybe[1], Florian Rämisch[1], and Tracy Hoffmann[1][0000−0001−8718−9536]

University Library Leipzig, Germany
{konstantin.freybe,florian.raemisch,tracy.hoffmann}@uni-leipzig.de

Abstract. In this paper we reflect on our research of Japanese video game culture, with focus on strategies of interdisciplinary collaboration. We understand our collaborative research as ongoing negotiation that aims at finding common ground between researchers from different backgrounds. We decided not to work on a single extensive question over the research period. Instead, we chose to work on a number of smaller problems (called Tiny Use Case (TUC)) that are aligned with superordinate research interests. Various methods from both humanities and information sciences were adapted and customized to these needs. A fundamental mutual understanding is essential for the various tasks in the team. The right choice and mix of methods and tools does not only depend on the specific team constellation (age, backgrounds, skills) but also a matter of available resources such as time, and the flexibility to explore.

Keywords: Interdisciplinary Collaboration · Mixed Methods · Tools · Software Development · Game Studies.

1 Introduction

In this study we reflect on our research, focusing on strategies of interdisciplinary collaboration. This is not only about working together, but sharing knowledge and establishing a mutual understanding. Our pursuit of this goal will be contextualized within our current research of Japanese video games.

We found that several adjustments to pre-existing concepts were beneficial to our work. We present our strategies in more detail in later sections, but sending ahead a brief summary hopefully helps drawing the connection from research content to our collaborative strategies more easily.

We understand our collaborative research as ongoing negotiation that aims at finding common ground between researchers from different backgrounds. Flexibility is crucial for collaboration, as long as it means to balance freedom of action with bindingness of reached agreements. We formalized this in what we call *TUC* workflow which is described after a brief introduction of the project diggr. The next section provides the evolution of the research interest and points to different methods we used to collaborate. These methods are presented in the following section. After a reflection about limitations of our approach we will end this paper with a conclusion about our work.

2 The diggr Project

diggr is a research project funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and conducted by the IT department of University Library Leipzig and the Institute for Japanese Studies of Leipzig University. Our research focuses on Japanese video games in the context of global resp. globalized video game culture. Six members of both the library and the Institute for Japanese Studies with different disciplinary backgrounds (Information Science, Librarianship, Cultural Studies, Japanese Studies) will attempt to integrate expertise in data management directly in the research of humanities scholars from 2017 until 2019. According to Tabak [11], during the project, the researchers take one of two roles as a humanities scholar (H) or digital expert (D) or a combination (DH). "The H-role provides the content and D-role deals with the technical aspects of DH projects." [11]

The project pursues two goals. Firstly, build a data driven research infrastructure that uses e.g. Linked Open Data technologies and provides scholars with best practice solutions. Secondly, generate substantial contributions to video games research in general, research of Japanese video games more specifically. Our two humanities scholars lead their own sub-projects. The other staff members pursue tasks like software development, system administration or data modeling. One of these sub-projects will provide the context for our discussion and should, therefore, be presented as well. But before doing so, we discuss our adjusted use case structure which both sub-projects follow.

3 Tiny Use Cases

The beginning of the project presented itself as a challenge, as both the data situation and the required technologies were unclear, which in turn made it difficult to formulate objectives [6]. Therefore, the development of workflows for joint research was an important first step in this project. In order to be able to test them in research practice right away, we decided not to work on a single extensive question over the research period. Instead, we chose to work on a number of smaller projects (called Tiny Use Case (TUC)) that are aligned with the superordinate research interests. With their help, explorative approaches could be developed which promoted collaboration between information technology and content-oriented researchers. TUCs are designed to be conducted in rather narrow time frames of approximately three to four months. Generally speaking, a TUC is structured as shown in Figure 1 and as follows:

Mediation of the research interest/object H attempt to convey their research interest for each TUC as a research question to the team. This is as sensitive as it is critical. Without at least a basic understanding of the research interest presented to them, D cannot be expected to provide guidance in regards to software solutions that fit H's requirements. In turn, H have to adapt to the perspective of D if they want to be able to evaluate whether a proposed software

does actually solve the task. This leads us to the next step, where negotiation shifts from conveying an idea to assessing adequacy of software tools.

Exploring software solutions In order to enable H to specify their software requirements appropriately, knowledge exchange is key. The basic idea of our approach is that D and H educate each other about the respective domain specific blind spots which leads to a common understanding and a shared technical terminology.

Evaluation A TUC workflow usually concludes with an evaluation. This is not only helpful for tracking progress of the team's work. It also allows us to critically reflect on the research conducted and to determine whether we reached the goals we set for ourselves. By evaluating frequently, as opposed to a single evaluation towards the end of a project, we have opportunity to thoroughly document our work.

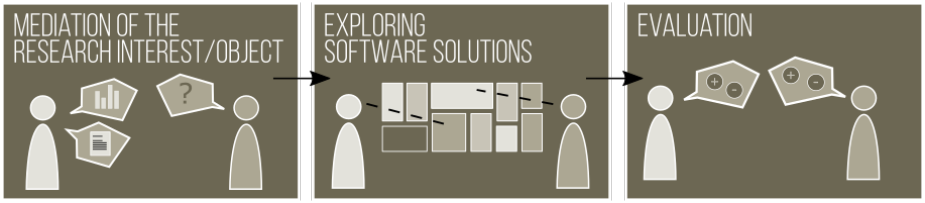

Fig. 1. TUC Workflow

4 Subproject: Video games as Practice – Culture as Negotiation

One researcher in our team is investigating video game fan practices on social media in order to learn about their position and scope of action within the gaming industry. Many video game related practices on social media services like YouTube translate consumption practices into public or semi-public performances. By reconstructing careers of YouTubers, their transition from users to content creators to Influencers[1], H1[2] intends to learn about the relation between fan practices, consumption and labour. What kind of labour is it being a *Youtuber*, which tensions do they face and how do attempt which changes in the gaming industry?

[1] Influencer Marketing is a name for a strategy pursued by advertisers. The presumably stable relation between YouTubers and their audiences is targeted in order to convey advertising messages.

[2] H1 refers to a single H-researcher

4.1 Tiny Use Case 1

H1 chose the video game series *Metal Gear* as a topical frame. In TUC1, our first and rather prototypical TUC, H1 was interested in the reconstruction of canonicity among the 31 titles that are associated with the series. He began his exploration *in* one of the games [8] and suspected that the design of a particular mission is addressing a specific, hardcore fan audience. H1 encountered difficulties explaining the connection between his gameplay findings, canonicity and how databases could be helpful. Our computer scientists found it hard to understand H1's goal from a verbal report of the in-game situation alone. In order to overcome this obstruction, H1 adopted a popular practice on YouTube, and recorded a playthrough of the aforementioned mission, including audio commentary. This video was then uploaded and shared via YouTube.[3]

In abstract terms, H1 found that the design of the *Deja vu* mission showed various references to titles from the Metal Gear series. Regarding gameplay mechanics (i.e. the design of player interactions with the game software), H1 found an odd, normative distinction in how players could interact with these in-game references. While being an optional objective, successful manipulation of the game world is rewarded with a personal message from lead designer and producer Hideo Kojima.

Equipped with a list of titles that were referenced in the aforementioned gameplay situation, we investigated Kojima's contribution to the productions. We approached this by analyzing credit information from transcripts of staff rolls[4]. This allowed us to associate person names with functions and production units, and to count the roles attributed to individual names. H1 interpreted his in-game findings as an attempt to draw an image of Kojima as authorial figure and to enforce the canonic status of a subgroup of games from the *Metal Gear* series. While we could reconstruct that Kojima had the highest number of role attributions, several other staff member executed multiple production roles as well, indicating an inner circle surrounding Hideo Kojima. This dismantles the notion of him being the single author and shows that video game productions are predominantly team efforts. This raises questions on why and how the prominence of Hideo Kojima is maintained.

4.2 Tiny Use Case 3

When it was time to begin TUC3, the focus then shifted to YouTube. This platform was chosen from various social media services. Aside from YouTube, we considered Twitch, Facebook, Twitter and Patreon. But eventually we chose to limit ourselves to researching YouTube due to its relatively high API request quotas and accessibility. This time, the research question was based on the assumption that practices like *Let's Plays* closely relate to consumption practices.

[3] METAL GEAR SOLID V : GROUND ZEROES - Kojimas Kanonisierungsstrategie? https://www.youtube.com/watch?v=Z1frZ-zWptM

[4] This is quite similar to movie credits.

If public consumption of video games generates considerable income for the respective person, is it reasonable to view this as labour?

TUC3 was divided in three episodes, labeled with *a*, *b* and *c*. 3a was designed to perform an assessment of the field, YouTube in this case. H1 stated that he views the relation between YouTubers and their audiences or communities to be of critical importance for social practices on YouTube. Otherwise, efforts like e.g. *Influencer Marketing* would make little sense and could not be as popular and effective. The focus on social interaction determined the kind of data that had to be procured.

TUC3a produced valuable orientation and knowledge on how to cater to the researcher's requirements. The next two phases were headed in a similar direction: 1. extend the subject area, or: how much more data from YouTube can be handled with what tools?, and 2. stabilize the more advanced prototypes.

In order to decide on appropriate software solutions, D commissioned H1 to try already existing tools and document this as user stories. These documents provided the foundation for D to extrapolate H1's requirements. This includes tasks like interface design.

Although diggr strives for re-using existing solutions, this is not always possible. In order to enable individual researchers to deal with millions of comments on their own, semi-automated means of analysis seemed very appealing to us. DH-Tandems were formed in order to educate H1 on basic functionalities and align these methods with his superordinate methodological considerations.

Frequent evaluation of our progress did enable us to negotiate viable solutions. The results of these discussions were documented by H1 as requirement profiles. To experienced software developers, a requirement profile might be a rather common tool. For a humanities scholar who is somewhat distant from IT matters, this might seem rather unfamiliar and by no means self-explanatory.

5 Inventory of Tools and Methods

In this section we describe the methods and tools we used in the TUCs. We made use of various methods from both humanities and information sciences and adapted them to our needs. "As research generally is a creative process, some of the most interesting research questions only develop over time" [10], it demands of us to "welcome changing requirements" [1]. The methods presented here are ordered upon their possible introduction into our workflow, but can also occur at later or earlier stages depending on the course of the TUC.

Figure 2 illustrates the process that lend to our inventory, beginning with selecting and customizing methods from rather specific domains. If an evaluation shows that chosen methods sufficiently helped solve their task, they are introduced to our inventory, which we discuss in the following sections.

5.1 Research Let's Play

Starting diggr's search for traces of Japanese video games turned out to be difficult, at least difficult to explain. Is there a shortcut that bridges the gap

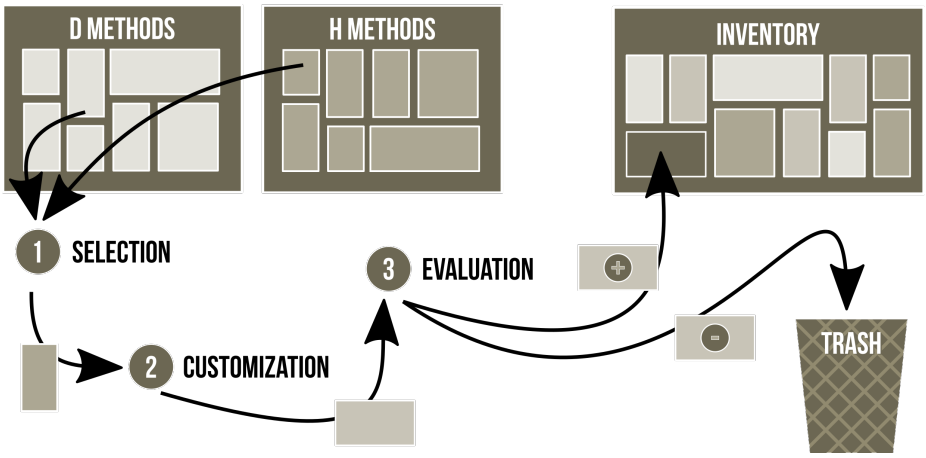

Fig. 2. Inventory Flow

between H1, who spent many hours with the games, and D, who might only know the titles from trailers or hearsay? How can in-game research be presented to team members so they provide a starting point for further research? H1 adapted a common practice of video game enthusiasts called *Let's Play*. While videos of this genre may be very diverse, they all have in common that gameplay is recorded as video and then uploaded online for others to view. Our team benefited from the low threshold and relative immediacy of this approach and, accordingly, we could soon start working on solutions. As the *Research Let's Play* can be seen as a protocol of an action or event, it is to be preferred over a live demonstration as it can be used as reference at a later time.

5.2 User Stories

Finding a common language is essential, when it comes to accurate description of expectations on functionality and user interface of software tools. D tend to overestimate the IT skills of H, while solid knowledge of the current research is expected. On the other hand, H have to be careful not assume that D are aware of Humanities specific presuppositions. Simultaneously, H depend on D's assessments regarding practicability of technical solutions.

In our group user stories turned out to be a good method to perform requirements engineering. H describes his expectations and wishes on the tools to be built, in simple and clear sentences. No attention is paid to the actual or presumed effort necessary to realize these. This is essential as some features appear simple to implement but aren't and vice versa. The user stories are then discussed in a feedback round in the group. D and H together agree on the desired and realizable feature set.

However, these procedures might yield the insight that some requirements cannot be met. In our case, this occurred when researching various social com-

munication services in regards to their API accessibility. We discussed Facebook, Twitch, Twitter, YouTube and Patreon. Restrictive APIs and issues of privacy protection were reasons to focus on YouTube. The decision on a subject area therefore is the result of frequent adjustments and ongoing negotiation between the involved parties within the team. Collaboration has to be intensified in order to develop solutions when requirements can neither be fully met nor abandoned. In the following section, we present one way that helped us solving tasks that require a stronger focus.

5.3 DH Tandem

The DH Tandem consist of two persons: D and H. It is a temporary working group committed to a specific and small work package, while the rest of the team works on other topics. In our case, the Tandem worked on the design of the research dataset.

In the process both, H and D become domain experts, as H gets a better understanding of the process of acquisition and composition of data, while D is introduced to the scope of the research question. During the tandem sessions, both parties develop a common vocabulary, which does not consist of new words, but technical terms from the fields of both D and H. The tandem members function as translators and contact persons outside the tandem for the rest of the team. They can be seen as the collective product owners of a research question. Therefore in the tandem sessions, both parties, D and H, are on eye-level, but with a clear understanding of their roles. Therefore, it is possible for them to negotiate on the requirements of the software and research dataset. While cost here usually is not the limiting factor, it is human resources and temporal constraints that pose limitations on the final feature set and extent.

5.4 User Interface Design

Traditionally, when it comes to software, magic is allowed to happen – even desired – to surprise the customer and enhance its user experience. Capability, usability, performance, reliability, installability, maintainability and documentation appear to be the only metrics to be accounted for when evaluating customer satisfaction with software products [7].

In contrast, H do not strive for surprises, but comprehensibility. Magic is not allowed. For reproducible science, the flow of data in the program, the intermediate steps, manipulation, modifications, enrichments etc. need to be made transparent to H. Lack of transparency of the methods and transformations applied appears to be a common problem with research software, as Gibbs et al. [5] point out.

Burghard et al. [3] identified two main problems in the design of linguistic annotation tools: Wrong or counter intuitive feedback provided by the User Interface as well as unconventional controls. Both cases never occurred in our

software projects so far, as the control elements as well as feedback by the software are designed in close cooperation between D and H. Requirements changes were negotiated in the DH-Tandem.

5.5 Provenance

In contrast to commercial software development, scientific development requires transparency and comprehensibility in regards to solutions and results. While individual TUCs are designed to be of short duration and intentionally limited scope, they line up iterative steps to contribute to superordinated research interests. Accordingly, this means including previously used data. In terms of comprehensibility, this iterative progression leads to the requirement to trace back how data was used and manipulated by whom. This is important for assessment and critique of the methods used and in extension the research conducted.

To allow for comprehensibility under these premises, this metadata about the origin and modifications of a file or data set is stored alongside the files. It is shipped with every research dataset in our group. To ease creation and use of provenance data the diggr team developed a tool for the creation, modification, display and export of provenance information: *provit*. [9] The tool is designed to be used in small research groups or by individuals. It aims to make retrieval and creation of provenance information as easy as possible.

5.6 Graphical User Interfaces

Graphical User Interfaces (GUIs) are essential for accessible research infrastructures. Lower entry bars, allow more scholars to work on their research questions empirically [10]. The reasons why matters of interface design are important for H are twofold. Firstly, H are commonly not accustomed to navigating through code or command line interfaces. Interfacing via frontend is less time consuming for H than to learn the required skills for pursuing alternatives. Secondly, H might lack the experience in operating software code and therefore may be unaware of risks. In order to avoid setbacks in the form of accidental deleting of files or causing fatal errors, interface design also becomes a precaution.

In contrast usable GUIs require a lot of effort in creation and maintenance. Insisting on GUIs for every experiment in the research process conflicts with the fast paced creative research process in general. (ib.) Often the tools D develop are only used once, or for a very limited amount of time. To verify our intuition, we developed GUIs for two applications: Human verification of automatically created links of entities of different databases and research of missing information on Japanese video game companies. After the tasks were completed, it was concluded, that the effort required to develop and maintain these tools was to much compared to their utility.

Not developing GUIs is not an alternative we could afford, so we decided to use parameterized GUIs, as they provide easy to use graphical representation of the data, with no need to navigate through code or command line interfaces. The advantage is, that instead of building hardly recyclable specialized GUI we

now combine existing widgets, spend less time coding and more time educating H how to use the tools. Collaborating in this way was perceived to be way more sustainable, as the skills H learns in here are also useful outside the limited scope of a TUC, which cannot be stated for specialized GUIs. We acknowledge that other projects might come to different solutions for their projects, as our workflow is quite special and fast paced.

Elasticsearch, Logstash, Kibana (ELK) represents one of the most heavily used software stack in data science. Its popularity comes from its combination of powerfulness and ease of use. While Elasticsearch is a very powerful search engine, Logstash is a data processing pipeline collecting and aggregating data. Kibana is a frontend which can be used by the end user to operate Elasticsearch. With its integrated filtering and graphing tools it is useful to explore new datasets. ELK's great benefits for H lie in emphasis on data exploration and dynamic visualizations. In our case, H require a solution that allowed for exportable visualization that maintain the connection to the referenced or aggregated texts. Since H intend to employ various software tools for different purposes (preselection of data, orientation, visualizations outside of ELK), ELK proved very useful as hub where all research data is stored and from where derived datasets can be send to other tools for further processing.

Jupyter Notebooks are used by all team members at almost all stages of the software development process and research workflow prototyping process. "Jupyter notebooks are one means to make science more open." They "embody the FAIR (Findable, Accessible, Interoperable, Reusable) principles for digital objects and assess their utility as viable tools for scholarly communication." [2] In recent years they become popular for sharing research results with their underlying data and algorithms in one citeable research object. With this Jupyter Notebooks support the transparency and reproducibility of the research process.

Jupyter Notebook is a web based development environment and interactive user interface. The notebook server can be run in a data center. This turned out to be useful, as the computers H uses, sometimes are not powerful enough to run complex tasks in a reasonable amount of time. The notebooks are linear lists of cells, where each cell can be a piece of code, documentation, formulas, tables, images, plots and videos. With that, it can even be used to create interactive collages or even publications. The cells are executed one after another. Changes in one cell do not require the other cells to be rerun. E.g. to train a complex machine learning model, and then prototype the further data processing pipeline is very easy. This feature makes it a great tool for prototyping workflows and experimenting with datasets.

Jupyter notebooks also can be used in combination with repositories such as Github and Zenodo (for versioning and publishing). They offer a great intermediate step between providing a Graphical User Interface and exposing the scientists to a Command Line Interface. Getting in touch with source code in an environment which, through enrichment with pictures, explanations, plots and instructions can be way more appealing to a novice than a classical Integrated

Development Environment (IDE). Technical details are not hidden away, which makes the whole workflow more transparent for the whole research group [10].

From H's perspective, this is a valuable solution because H's actions are limited to small customizations at specific locations in the source code, as opposed to full access. The risk of causing damage to the software by unskilled editing or other reasons can be quite burdensome for H. Being freed of the need for permanent caution, H can contribute to the research process safely while also benefiting from working software – even at prototypical development stages. This is especially useful when refering to data analysis and (quasi-) dynamic visualizations.

Yet, being able to understand a scripting language is a valuable skill for H (ib.). We learned, that it helps H to follow the software development process more closely and get a better understanding of overall mindset.

5.7 YAML as configuration language

H needs to be able to configure programs according to their research interests and requirements, without having to build a GUI. A comparison of different machine data formats and data serialization formats led to the conclusion that *YAML Ain't Markup Language* (YAML), a data serialization format, is easy to be written by H and to be processed by our tools. YAML is with a clear syntax which shortens the time required by H to learn how to use it within the project.

In contrast to Jupyter Notebooks, which are used for analysis purposes, maintenance of configuration files turned out to be more practical when it comes to data acquisition. When assembling a dataset, e.g. from selection of YouTube channels, maintaining and managing YAML configuration files are relatively simple tasks which H can learn to solve more quickly than it takes D to write and provide Jupyter Notebooks – not to mention fully fledged GUIs.

5.8 Markdown as markup language

For texts which are not to be printed, like README files and documentation, blog posts, text drafts, meeting protocols, etc. document oriented file formats like Open Document Format and Office Open XML appeared unsuitable, as we often ran into formatting and paging issues. While H mostly used Microsoft Word and other WYSIWYG text processors, the information scientists preferred LaTex. As a compromise we decided to use Markdown as markup language for text in general. This has the advantage, that it can be written in a collaborative manner with CodiMD, and the results are easily and predictably convertible to Redmine Markup (for the issue tracker), HTML (for blog posts), PDF (for documents) latex (for publications) etc.

While having many output options, the clear and minimal syntax make it easy to read and write. It can be used directly within Jupyter Notebooks, or semi-WYSIWYG editors like CodiMD. There is immediate feedback on the correctness of the markup used, which helps both D and H to learn and remember the new languages.

Using Markdown and YAML has proven to be an effective approach for H and D to collaboratively work on projects with the same tools. This helps both the H and the D to better assess the skills and expectations of each other.

6 Limitations

The methods presented here have proven to be effective within our research group. Yet, they are far from being recommendable as best practices. The characters in the team and our environment aid intense collaboration. The relatively low demographic diversity in our team (all members between 28 and 38) may have contributed to finding common ground and a shared terminology. Team building events, such as gaming sessions in the GamesLab of the University Library Leipzig helped to build our team and increase the understanding for each other and the research context.

Spatial conditions might have made some methods and approaches more favorable than others. The whole team shares an office, which is (almost) exclusively used by diggr. Therefore face to face communication is common and problems often can be solved immediately without using an issue tracker.

Our TUC workflow has proven beneficial for collaboration in our team. But the applicability of methods outlined in this study is likely to depend on further customizations and adaptions by the adopting teams. A continuous evaluation of the research process helps to find compatible methods.

7 Conclusion

In this paper we presented our experiences, methods and tools for collaboration inside the digital humanities project diggr. A fundamental mutual understanding is essential for the various tasks in the team. To choose the right mix of methods and tools for the specific team constellation is a challenge and requires time, flexibility and the willingness to experiment.

It has shown that the humanities can benefit from agile approaches and methods in computer science. Computer science can also be enriched by the humanities, for instance through the continuous reflection of one's own work [4] and the need to make each step transparent and reproducible.

With this paper we hopefully contribute to further studies about the collaboration in digital humanities projects.

References

1. Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R.C., Mellor, S., Schwaber, K., Sutherland, J., Thomas, D.: Manifesto for agile software development (2001), http://www.agilemanifesto.org/

2. Boscoe, B.M., Pasquetto, I.V., Golshan, M.S., Borgman, C.L.: Using the jupyter notebook as a tool for open science: An empirical study. CoRR **abs/1804.05492** (2018), http://arxiv.org/abs/1804.05492
3. Burghardt, M.: Annotationsergonomie: Design-empfehlungen für linguistische annotationswerkzeuge. Information - Wissenschaft & Praxis, vol. 63 (2012). https://doi.org/10.1515/iwp-2012-0067, https://www.degruyter.com/view/j/iwp.2012.63.issue-5/iwp-2012-0067/iwp-2012-0067.xml
4. Freybe, K., Hoffmann, T.: Iterative bearbeitung von forschungsfragen. In: Burghardt, M., Müller-Birn, C. (eds.) INF-DH-2018. Gesellschaft für Informatik e.V (2018). https://doi.org/10.18420/INFDH2018-04
5. Gibbs, F., Owens, T.: Building better digital humanities tools. Digital Humanities Quarterly **6**(2) (2012), http://digitalhumanities.org:8081/dhq/vol/6/2/000136/000136.html
6. Hoffmann, T., Freybe, K., Mühleder, P.: Workflows zur datenbasierten videospielforschung - am beispiel der populären videospielserie metal gear solid. In: Eibl, M., Gaedke, M. (eds.) INFORMATIK 2017. pp. 1113–1124. Gesellschaft für Informatik, Bonn (2017). https://doi.org/10.18420/in2017_113
7. Kekre, S., Krishnan, M.S., Srinivasan, K.: Drivers of customer satisfaction for software products: implications for design and service support. Management science **41**(9), 1456–1470 (1995)
8. Konami Digital Entertainment: Metal Gear Solid V: Ground Zeroes (Mar 2014), Sony PlayStation 4
9. Mühleder, P., Rämisch, F.: provit (Dec 2018). https://doi.org/10.5281/zenodo.2268521, https://github.com/diggr/provit
10. Reiter, N., Kuhn, J., Willand, M.: To gui or not to gui? In: Eibl, M., Gaedke, M. (eds.) INFORMATIK 2017. pp. 1179–1184. Gesellschaft für Informatik, Bonn (2017). https://doi.org/10.18420/in2017_119
11. Tabak, E.: A hybrid model for managing dh projects. Digital Humanities Quarterly **11**(1) (2017), http://www.digitalhumanities.org/dhq/vol/11/1/000284/000284.html

Combining hermeneutic and computer based methods for investigating reliability of historical texts

Alptug Güney[1] and Cristina Vertan[1] and Walther v. Hahn[1]

[1] University of Hamburg, Hamburg Germany
{cristina.vertan,alptug.gueney}@uni-hamburg.de,
vhahn@informatik.uni-hamburg.de

Abstract. Within the framework of the project HerCoRe[1] we are analyzing two historical works from 18th century "History of Rise and Decay of the Otoman Empire" and "Description of Moldavia" (both written by Dimitrie Cantemir) and investigate them with regard to the historiography of its time. We evaluate the usage of sources by the author and also the reliability of his references. We also seek to shed more light on the motivation behind the writing-process of these works by taking into account the political and cultural dynamics of the time and the position of Cantemir within the Ottoman elite. To determine missing or incorrectly translated parts of the work, the German and English translations are also compared with a copy of the Latin manuscripts. This comparative approach serves also to discuss the causes of the (un)conscious mistakes and omissions in the translations. We are performing this study by means of hermeneutic and IT approaches.

Keywords: historical documents, uncertainty and vagueness annotation, hermeneutics.

1. Rationale of the research

Dimitrie Cantemir (1673-1623) was prince of Moldavia (a historical area including regions from current eastern Romania, Republic of Moldavia and some parts from Ukraine), He was a man of letters, philosopher, historian, musicologist, linguist, ethnographer and geographer. He received education in classical studies (Greek and Latin in his country of origin), then he lived for several years in Istanbul where he learned Turkish, and familiarized himself with the cultural traditions of the Ottomans, met important persons around the sultan and learned a lot about the history of the Empire. After a very short period of being prince of Moldavia he was forced to immigrate to Russia, where he became an important person at the court of Tsar Peter the Great. During this period, his works gained attention in the Western countries. He became member of the Royal Academy in Berlin and, on

[1] Research described in this article is supported by HerCoRe project, funded by Volkswagen Foundation (Project no. 91970)

their request, he produced the two books which are the target of this proposal:

- *Descriptio antiqui et hodierni status Moldaviae*, written in Latin, a history of his country in which he describes not only pure historical facts but also traditions, the language, as well as the political and administration system. Local denominations and toponyms, as well as names are written in Romanian with Latin script as his intention was to demonstrate the Latin origin of his folk. The transcriptions are not standardized and one retrieves for the same toponyms, several name variations. Quotations as known today were very rare, there is no bibliography. According to [3], as there was practically no consistent previous work about the region, Cantemir himself was not particularly careful with indicating sources of knowledge. The work is accompanied by a map, the first detailed cartography of the region. The names on the map are in Romanian language. The Latin original was translated for the first time into German, and only later - at the middle of the XIXth century - into Romanian. The Latin manuscript seemed to be lost for a long time, so that the first Romanian translation was following the German one. The German translation is containing editorial notes of the translator.
- *Historia incrementorum atque decrementorum Aulae Othomanicae*, the history of the Ottoman Empire. In contrast to the previous work about Moldavia, here Cantemir indicates very carefully the sources of information. [3] supposes the existence of previous works, known in the western countries, behind this decision. This work was written also on the request of the Academy in Berlin. Cantemir follows the same principle: text in Latin, while the toponyms and local denominations are written this time in Ottoman Turkish. Although there were already some previous works about the Ottoman Empire, the novelty of his approach is the quotation of Turkish sources. The reliability of these sources is untrusted sometimes by Cantemir himself. The original manuscript (or a copy of it) reaches the western world after Cantemir's death, carried by his son to London. Here, a first translation into English is produced: The history of Raise and Decay of the Ottoman Empire. The translator reinterprets the texts, probably also being confused by the presence of Turkish information sources, which at that time were perceived as completely unreliable. The Latin original remains lost for centuries and is rediscovered only at the end of the XXth century in the USA. Thus, the German translation is based on the English one and inherits the same alterations, and presumably adds new ones. The Romanian translations, in contrast, use the Latin versions. The last translation [2] is being used in this research.

Until now there is no systematic study on the reliability of the text sources in Cantemir's works, nor the degree of alterations produced by the translations of the two works.

Given the fact that both works became standard reference for western authors until the middle of XIXth century, it is expected that their reception influenced also following historical material. There is no reprint/new edition of his works in German or English. There are, however, several reprints of the Romanian versions. Recent Romanian translations of Decriptio Moldaviae are done after the original Latin manuscript.

A lot of works were dedicated to the personality of Dimitrie Cantemir and its perception in different parts of Europe. A study of the reliability and consistency of the historical facts (as they are described in the latin copies) and their translations is practically impossible ~~to be done~~ only with traditional hermeneutic methods. One needs expertise at the same time in Latin, German, English, Romanian, Turkish, to enumerate just the main languages used in the two books, which additionally sum up to a quantity of about 1000 pages. Both German editions are printed in "Fracture" ("black letter") script, which nowadays is very difficult to be read.

Already in the 1920s it was demonstrated (by using only a selections of texts), that the translations are not respecting the original all the time. E.g. information sources indicated by Cantemir were omitted, because they seemed too unreliable to the translator.

In the XXth century researchers claimed that some of the sources, persons and facts quoted by Cantemir were not existing at all (e.g. [1]).

But given the:

- geographic distribution of material (originals in libraries in USA and Russia; translations and copies all across Europe; most part of the quoted sources in Turkey),
- the multilingual character of the materials to be investigated (Latin, German, Romanian, English, Turkish at least) and
- The Quantity of data which has to be processed in parallel,

no study about the reliability and consistency of the original and the translations could have been performed until now.

In the HerCoRe project we propose a mix of hermeneutic and IT-methods in order to:

- compare the Latin copies and the English and German translations,

- identify translation mistakes or gaps (made by purpose or not),
- search after the quoted works and identifiy related Ottoman sources,
- analyse Cantemir‘s writing and discourse style,
- assess the importance of the work in the Ottoman studies and compare them with other works contemporaneous to Cantemir or follow-up research about the Ottomans,
- develop electronic resources which may be of use for follow-up work about the Ottoman empire and the history of Balkans.

2. Hermeneutic investigation

The hermeneutic investigation concentrates on the identification of sources quoted directly or indirectly by Dimitrie Cantemir, as well as the mentioned places, persons, events and dates.

The two works are very different with respect to the quotation style. While in the “Description of Moldavia” the quotation sources are almost missing, in the “History of rise and decay of Ottoman Empire” the author refers explicitely to different sources. However, there is no quotation style like in modern scientific works. Most references are real quotations or the author indicates the source of quotation through syntactic phrases followed by a reformulation of the semantic substance of a text section. Especially these cases are subject to the hermeneutic investigation.

By now we identified the main works quoted by Cantemir. These works are available only in paper form and are written in Ottoman Turkish (with Arabic alphabet) thus only a manual comparison can be performed.

This systematic comparison led to a very unexpected result: we observed that linguistic expressions of certitude (e.g. *“for sure”*, *“without any doubt”*) are not an unambiguous indicator of the reliability of the quotation. E.g.: Cantemir is sure that all investigated sources mention 4 sons of Sultan Bayazid. However, all reliable sources of the time mention that the sultan had five sons.

We do not know why these inconsistencies occur. One possible motivation is the context in which he wrote the two books: in exile in St. Petersburg, probably with few notes at hand, that he made in Istanbul. Whilst we cannot find the reason of the inconsistency, the hermeneutic analysis showed that a pure automatic annotation (searching for quotation marks) will not help in the case mentioned above, as the semantics of the quotation mark does not match the degree of reliability of the quoted information. This is something which cannot even be inferred by automatic methods; at least

not at this stage, where documents in Ottoman Turkish are rarely digitised and no linguistic tools are available for this historical language variant.

A second part of the hermeneutic investigation concerns the collection of persons, places, and domain specific concepts which are mentioned by the author. An automatic identification is practically impossible, as names are not standardised (e.g. for the city of Iasi in Moldavia we identified at least 12 writing variants).

The third part of the hermeneutic investigation is concerned with the identification of missing paragraphs in the German and English translation. First result: all paragraphs written in the Latin original with Arab characters were systematically omitted. This leads to misunderstandings in the two translations.

3. Computer-based approach

Digital methods can facilitate analysis on the reliability of translations but also of the historical facts claimed by the author [8]. In order to be effective, these methods must consider an intrinsic feature of all natural languages: the ability of producing and understanding vague utterances. The project HerCoRe aims at modelling and annotating five levels of vague assertions

1. the text uncertainty (uncertain readings, losses, translations, multilinguality, etc.),
2. the linguistic vagueness (metonymies, vague adjectives, comparatives, non-intersectives, hedges, homonyms,),
3. the author reliability (genres, time style, contemporary knowledge),
4. the factual uncertainty (range expressions, time expressions, geo relations), and
5. historical change (named entities, abbreviations, meaning changes)

We develop an annotation formalism which allows for:

- the mark-up of different types of vagueness and its source; the implementation of a set of inference rules for the combination of such vague features to calculate an overall result of their reliability;
- the definition of a similarity measurement of the inferred results obtained for the same queries on different translations. The system architecture is presented in figure 1. It relies on annotation on 4 levels (linguistics, lexical markers for vagueness/uncertainty, ontological and factual/quotation markers).

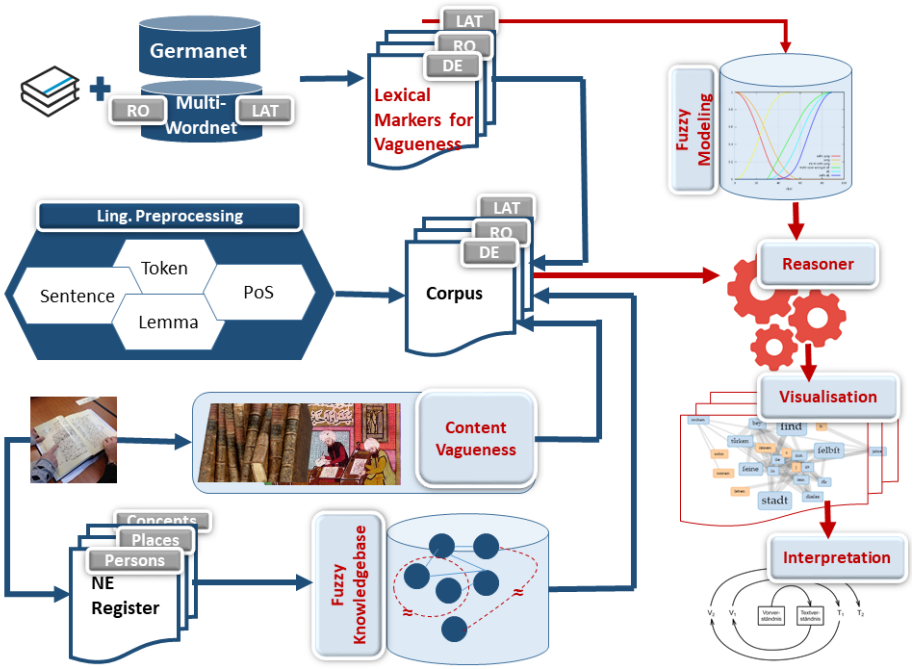

Figure 1. HerCoRe System Architecture

For the detection of linguistic vagueness we follow a multilingual approach. We collected the above listed indicators in the three languages involved in the project (Latin, German and Romanian). Based on [5] [6] we distinguish between:

- Vague quantifiers, e.g.: some, most of, a few, about, etc.
- Modal adverbs, e.g.: probably, possibly, etc.
- Verbs e.g.: to believe, think, prefer, assume etc.
- Lexical quotation markers , e.g. introduced by quotation marks or verbs with explicit meaning (say, write, mention),
- Inexact measures and cardinals.
- Complex quantifiers
- Non-intersective adjectives
- Implicit syntactic clues: mainly verb moods such as conditional-optative for Romanian, conjunctive mood or past perfect/pluperfect for Latin, all of them indicating a “counterfactive” or non-reality (doubt, hear-say, possibility, etc.)

The initial collections of linguistic indicators are enriched through synsets in the corresponding Wordnets.

The knowledge base backbone is ensured by a fuzzy ontology modelled in OWL2. We distinguish between fixed concepts and relations (like geographical elements: river, mountain, island) and notions for which several "contexts can be defined. E.g. a geographical notion like "Danube" is within one historical context a border of the administrative notion "Ottoman empire", and in another one the border to the so called administrative notion "Roman empire". The historical contexts are specified by further fuzzy data properties (e.g. time, placement).

4. A case study

In "The History of the Growth and Decay of the Ottoman Empire (1734)" Cantemir tells the story of a battle between the *Moldavian Prince Ștefan* and the *Ottoman Sultan Bayezid I.* Cantemir does not give the exact date of the battle, but one could think that this can be inferred from other details of the text. *Ș*tefan attacked the Ottoman camp in *Rasboeni*. According to the account of Cantemir, in this first confrontation, Ștefan lost the battle and retreated to his castle in *Neamț*. Then, after the inspiring speech of his mother in *Neamț*, he drew his soldiers together and stroke back the Ottoman army twice in succession. After the last defeat in Vaslui, Sultan Bayezid fled back to Edirne [7]. In the same paragraphs, Cantemir gives in a footnote some information - among many other detailed and important details - about the Moldavian Prince Ștefan, who fought two times against *Bayezid I.*:

"*He overthrew the renown'd Matthias King of Hungary, and wrested from him Transilvanian Alps […] His son Bogdan made Moldavia tributary to the Turks.*"

This account of Cantemirs includes several problems, which can mislead the reader and even a historian who does not have detailed knowledge about the Ottoman and Romanian (Wallachian and Moldavian) history.

Known and proven historical facts:

- There were two sultans with the name Bazeyid in the history of Ottoman Empire Bayezid I and Bayezid II but only the first one had the additional name "Yildirim" i.e. the Thunderbold. Cantemir mentions exactly this appellative and not the numbering (I or II) so we can exclude any typo or damaged spot in the manuscript. We checked this information in all translations and the Latin facsimile.
- The reign time of Bayezid I is known for sure (according to diplomatic text sources): 1389 – 1402.
- The frontiers of the empire at that time leaned already towards the Danube River and the Ottoman Empire was yet neighbor to Wallachia and Moldavia, thus a military confrontation with both principalities is historically possible.

- During this time Wallachia was ruled by several princes: Mircea I (1386-1395), Vlad I (1394-1397) and then again by Mircea I (1397-1418).
- The Ottoman chronicles report on a battle in 1391 between the Wallachian Prince *Mircea* and *Bayezid I* in a place called Arkaş (in Romanian *Rovine*). According to the Ottoman historians, Bayezid won the battle and Mircea recognized the Ottoman sovereignity [4].
- At the time of Bayezid Wallachia was ruled by Mircea I (1386-1395), Vlad I (1394-1397) and then again by Mircea I (1397-1418). In Moldavia there was just one Ruler called Ștefan (Ștefan I 1394-1399)
- There was a Moldavian ruler Ștefan III (1457-1512) who defeated the Ottomans in 1475 after a loss in Rasboieni, and who defeated also the Hungarian King Matthias (known as Matthias Corvinus [1443-1490]). Moreover, this ruler had a son called Bogdan who made Moldavia tributary to the Ottomans. These facts are confirmed by Ottoman chronicles [4].

At a closer look there is a strong mismatch between Cantemir and all other established chronicles, but a historian would not know here how to interpret the text:

- Is it referring to a battle against Moldavia or Wallachia?
- Which Ruler opposed Bayezid Yildirim?
- Where took the battle place?

An historian using only traditional methods (source inspection, reflection, re-evaluation of text) will face here a bunch of unsure and contradictory information, very difficult to resolve. An historian with less background knowledge about the Romanian history will be tempted to interpret wrongly the text section, either choosing the wrong rulers or the wrong place.

The HerCoRe System aims at helping historians in their interpretation, and suggests different reading paths. From the ontology and additional annotations the following inferences are possible:

- `Class Ruler and wasRulerof value 'OttomanEmpire' and has Name Bayezid and has Additional Name Yildirim -> Bayezid Yildirim 1389-1402`
- `Class Ruler and hasName 'Bayezid' and 'has additional-Name Yildirim' and hasBattlesIn some (Class Principality and belogsTo some (Historical Region and (liesIn value NorthDanube) -> Moldavia and Wallachia`
- `Class Ruler and was RullerOf exactly Moldavia and had BattleWith value 'Bayezid Yildirim' ->` this will show all rulers which fulfil the criteria.

Continuing this queries to the ontology, or invoking a complex inference chain the system will propose following solutions:

- The paragraph is about a battle in Wallachia in a place called Rovine and against a ruler *Mircea I*. -> contradicts Cantemir text: (it is not a Moldavian king but a Wallachian). The user may infer also that the place called '*Rasboe*' by Cantemir might be the place known as '*Rovine*'
- The paragraph is about a battle in Moldavia against the Ruler Ștefan I, and Cantemir mistakes the information about Ștefan I for Ștefan III
- The paragraph is about the battle in *Răsboieni* against the Moldavian prince *Ștefan III* and the mismatch here is about the Sultan (*Bayezid I* or *Bayezid II*). Thus, the battle place here called by Cantemir '*Rasboe*" would be '*Răsboieni*'.

All this information will have attached a score indicating a degree of truth. The latter one e.g. will have a lower score when introducing an additional scoring parameter: the metadata. The metadata will say, that the mentioned paragraph is within a chapter about the sultan Bayezid I, so it is less probable that Cantemir mistakes the name of the Sultan.

HerCoRe System does not aim at proposing a final solution. This decision is left entirely to the user/researcher, who is the hermeneutic subject and also can store the inference paths presented above as motivation for his choice.

5. Conclusions and further work

In this article we intend to show how hermeneutic and IT methods can be combined in order to investigate the reliability of historical texts (original and their translations). We show that a deep analysis can be performed only by the combination of the two approaches. Current research focus on the semi-automatic annotation and development of the ontology. Further work concerns the implementation of the reasoner and the visualisation of results.

6. Remarks on cooperation

In our project we need the cooperation between computer scientists, linguists (Latin, German and Romanian) as well as researchers in turcology. We cooperate also very close with the editor and translator into Romanian of the two works. The cooperation already revealed to date interesting aspects:

- The simple availability of raw texts in digital form does not help. E.g. the German translation of the History of Ottoman Empire is available from

the German Text Archive. However, the text was digitized for visualisation purposes. The usage of the underlying text versions leads to an unsorted mixture of paragraphs written by the Cantemir, his side notes and the comments of the translator. These parts, marked correspondingly in the TEI version are melted in the .TXT version. We had to invest additional work on separating these distinct parts.

- Mark-up in the editions have to be considered, as they can enrich the text. However, first one has to know the semantics of the mark-up.
- The ontological formalisation of the notions mentioned in the text was a great help for the humanist researchers leading to a better reflexion of the used notions.
- Many of the computational linguistics approaches had to be revised given the particularities of the historical text.

References

1. Babinger, Franz, 1927, Die Geschichtsschreiber der Osmanen und ihre Werke. Leipzig
2. Dimitrie Cantemir, Istoria măririi şi decăderii Curţii othmane, 2 volume, editarea textului latinesc şi aparatul critic Octavian Gordon, Florentina Nicolae, Monica Vasileanu, traducere din limba latină Ioana Costa, cuvânt înainte Eugen Simion, studiu introductiv Ştefan Lemny, Bucureşti, Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2015. ISBN 978-606-555-135-0 (978-606-555-136-7, 978-606-555)
3. Lemny, Stefan, 2010, Cantemirestii -Aventura europeana a unei familii princiare din secolul al XVIII-lea, Polirom Publishing House.
4. Parmaksızoğlu, İsmet (Ed.), Hoca Sadeddin, Tâc'üt-tevârih, Bd. III, Ankara, 1979, 153-158; Abdülkadir Özcan, „Boğdan", TDV İslam Ansiklopedisi, Bd. VI, 269.
5. Pinkal, Manfred, Semantische Vagheit: Phänomene und Theorien. In: Linguistische Berichte 70. 1980. 1-26. und 72. 1981. 1-26.
6. Pinkal, Manfred, 1985 Logik und Lexikon: Die Semantik des Unbestimmten.
7. Unat, Faik Reşit (Ed.), Mehmed Neşrî, Kitâb-ı Cihan-Nümâ, Bd. I, Ankara, 1949, 327; Parmaksızoğlu, İsmet (Ed.), Hoca Sadeddin, Tâc'üt-tevârih, Bd. I, Ankara, 1979, 200-201.
8. Vertan, Cristina and v. Hahn, Walther, 2014, Discovering and Explaining Knowledge in Historical Documents, In: Kristin Bjnadottir, Stewen Krauwer, Cristina Vertan and Martin Wyne (Eds.), Proceedings of the Workshop on "Language Technology for Historical Languages and Newspaper Archives" associated with LREC 2014, Reykjavik Mai 2014, p. 76-80.

It Takes a Village: Co-developing *VedaWeb*, a Digital Research Platform for Old Indo-Aryan Texts

Börge Kiss (Institute for Digital Humanities)[1], ✉ Daniel Kölligan (Dept. of Linguistics – Historical-Comparative Linguistics)[1[0000-0002-3134-8398]], Francisco Mondaca (Cologne Center for eHumanities)[1], Claes Neuefeind (Institute for Digital Humanities, Data Center for the Humanities)[1], Uta Reinöhl (Dept. of Linguistics – General Linguistics)[1, 2[0000-0003-2829-682X]] and Patrick Sahle (Cologne Center for eHumanities)[1[0000-0002-8648-2033]]

[1] University of Cologne, D-50923 Köln, Germany
[2] Johannes Gutenberg University Mainz, D-55099 Mainz, Germany

b.kiss@uni-koeln.de
✉ d.koelligan@uni-koeln.de
f.mondaca@uni-koeln.de
c.neuefeind@uni-koeln.de
uta.reinoehl@uni-koeln.de
sahle@uni-koeln.de

Presenters: Daniel Kölligan (humanities scholar), Claes Neuefeind (digital expert)

1 Research Goal: the "Humanities Problem"

The traditional linguistic approach in dealing with large text corpora includes the writing of concordances (or word indexes) to make co-occurrences of forms visible and to enable researchers to detect the frequency of usage patterns. The determination of meanings, functions, syntactic patterns etc. are mostly based on the individual assessment of a specialist drawing their conclusions from their intuition based on personal reading experience and more or less implicit knowledge of the texts (cf. e.g. [12, 20]). However, the more the quantity and quality of data increase, the more intractable these resources become for such traditional "tools" and approaches. At the same time, the recent wave of digitizing textual data makes the application of computational methods unavoidable in order to ensure the highest possible scientific standard in quantitative terms. The more information resources are made accessible, systematically prepared and annotated, the more urgent the need for their formal analysis becomes. The project described below is the interdisciplinary endeavour to prepare such a resource for the linguistic analysis of a large corpus of richly annotated textual data.

The objective of the *VedaWeb* project is to develop a web-based research platform for linguistic and philological work on Old Indo-Aryan texts. To this aim, multiple textual, lexical and other types of linguistic data are integrated in a single digital research environment. Various layers of texts and annotations are made searchable

and are linked with lexicographic information. At a later stage, the platform will offer personalized work spaces for researchers to enable collaborative work on the texts. Long-term sustainability of data and software will be ensured through cooperation with the Data Center for the Humanities at the University of Cologne (http://dch.phil-fak.uni-koeln.de). In order to build *VedaWeb*, a group of historical linguists, general linguists and computational linguists have teamed up with specialists in digital humanities at the University of Cologne. In addition, the project team collaborates with several international co-operation partners (e.g., from the University of Zurich and from Harvard University).

For illustration, concrete examples of humanities research questions that will be possible to address working with the *VedaWeb* platform include the following: In contrast to the rather intuitive description of meanings as described above, linking lexical units with and measuring their distance to all other units will allow to form networks of co-occurrences, enabling a more systematic and better-controlled semantic analysis. Linear distances on the one hand and co-occurrence frequencies on the other serve as variables for the configuration of these networks. This possibility will also further the study of lexical fields that so far had to be carried out "by hand" (cf. [9, 10, 17]). In the realm of morphology, given the data stored in *VedaWeb,* it becomes possible to study allomorphic variation within Vedic Sanskrit, e.g. the competing forms of the nominative plural masculine of stems in ending in *-a-* (e.g. *áśv-a-* 'horse'), viz. *-ās* (*áśvās* 'horses') and *-āsas* (*áśvāsas,* also 'horses'). With the help of *VedaWeb* these are currently studied in a project at the University of Cologne (http://ifl.phil-fak.uni-koeln.de/36486.html?&L=1), which tries to establish their synchronic distribution in the Rigveda according to various criteria such as agency and topic-hood and to propose hypotheses regarding their diachronic development.

2 Background and Starting Points

VedaWeb provides access to ancient Indo-Aryan texts written in Vedic Sanskrit, which are enriched with morphological and metric annotations as well as integrated with lexicographic information, making them searchable according to corpus-linguistic criteria. The pilot text of this project is the Rigveda, one of the oldest and most important texts of the Indo-European language family and the oldest extant one composed in Indo-Aryan, whose origin can be traced back to the late second millennium BC. With an extent of c. 160.000 words in nearly 40.000 lines of verse (comparable to the Homeric epics *Iliad* and *Odyssey* combined totalling c. 190.000 words), the Rigveda presents an extremely rich text corpus given its considerable ancestry. At a later stage of the project, further texts such as the *Atharvaveda* (c. 170.000 words), *Yajurveda* and Vedic prose texts such as the *Aitareya Brahmana* (c. 100.000 words) and the *Maitrayani Samhita* (c. 120.000 words) are also to be integrated into the *VedaWeb* platform to allow complex searches across several texts. Further possible features include user-specific annotation tiers, semantic searches, and the embedding of audio and video data of traditional recitals of the texts. The project aims to advance research in all areas of Vedic linguistics and philology, for example

in syntax (e.g. referential null objects [11], non-configurationality [15]), morphology (e.g. the Vedic *vr̥ kī*-type [21], *ya*-presents [13]), metrics [18] and word formation (e.g. compounds [19]).

The starting point of the project is a complete morphological annotation of the Rigveda, which was carried out at the University of Zurich over several years, where each word form has been annotated according to morphological categories (nouns: case, number, gender, verbs: tense, mood, number, person, voice, etc.). In addition, metrical information (provided by Kevin Ryan, Harvard University, http://www.meluhha.com/rv/) and syntactic information from different completed and ongoing research projects will be made available in *VedaWeb*. Metrical data for each word form note the frequency of attestation, the number of syllables, the metrical weight template (heavy and light syllables), any metrical variants (e.g. *indra* vs. *indara*) and the number of occurrences in different positions in the verse. Syntactic information comes from two sources: H. Hettrich (University of Würzburg) and O. Hellwig (Heinrich Heine University Düsseldorf) [8] have annotated all verbs and their arguments (e.g., direct objects) of the *Rigveda* as part of Hettrich's studies on the syntax of case in Vedic. This will allow searches for semantic categories and their morphological surface expressions (e.g. "Which morphological cases are used to denote instruments?"). Based on the Haig/Schnell annotation scheme GRAID (*Grammatical Relations and Animacy in Discourse* [7]), a project team lead by PI Reinöhl (http://sfb1252.uni-koeln.de/b03.html) is currently annotating selected Vedic texts for syntactic and pragmatic information which will allow searches according to types of referents (+/- human, +/- animate), clause types (e.g. main vs. relative clause), etc.

In the course of the project, multiple research and analysis tools are integrated into the platform. These include a combined search function for linguistic parameters (including lemma, word form, morphological and metric information), links for each word to entries in the standard dictionary of the Rigveda by Hermann Grassmann [3], the display of translations (including [2, 4, 6, 16]) as well as commentaries [14, 16], and the possibility of exporting sections of annotated text according to criteria selected by the user.

A key feature is to link the Rigveda with the C-SALT (*Cologne South Asian Languages and Texts*, http://c-salt.uni-koeln.de/) application programming interfaces (APIs) for Sanskrit Dictionaries (https://api.c-salt.uni-koeln.de/). Based on a TEI (Text Encoding Initiative) data model, the word forms in the text are linked to the respective lemmas in Grassmann's dictionary. In this way, it will also be possible to create links to lemmas in further Sanskrit dictionaries, for example to enable comparative, cross-dictionary searches (such as the DFG-funded *Nachtragswörterbuch des Sanskrit* at the University of Halle with whom a collaboration to this aim is planned, http://nws.uzi.uni-halle.de/). The direct linking of text and dictionary is a unique feature of the *VedaWeb* platform compared to existing resources of Old Indo-Aryan texts, e.g. the *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS, http://titus.uni-frankfurt.de).

3 Realization: the “Technical Solution”

The *VedaWeb* platform is implemented as a web application based on the Spring framework (https://spring.io). Spring serves as middleware providing controllers for data handling, service orchestration and different views on the data. Since the *VedaWeb* search should allow for linguistically motivated queries on different levels of annotations by means of complex, combined search criteria, a tailored search logic was implemented on the basis of the established search server Elasticsearch (https://www.elastic.co/). This solution provides fast search functionality and enrichment of search results (e.g. highlighting of lemmata vs. word forms). The frontend is implemented as a single-page application (SPA) based on React (https://reactjs.org) for building interactive front-end components. It makes use of Ant Design (https://ant.design/) for a clear and effective interface.

The TEI model adopted for *VedaWeb* plays an important role both with regard to the technical implementation as well as for data sustainability. As a cooperative project, the various sets of information to be integrated into *VedaWeb* have different sources and formats. For this reason, the various data resources had to be standardized in order to be employed in the *VedaWeb* application and made available for external projects. TEI provides a well-documented model and the necessary flexibility to achieve this goal, especially with regard to long-term sustainability.

The Rigveda has been modelled mainly according to an inline-markup paradigm: all of its text versions, annotations and translations are available in one TEI document. Versions are aligned on the level of the stanza while linguistic annotations are added to the single word forms of the chosen ‘Leittext’ via feature structures [22]. The exception to this markup-style are the references to dictionaries, which are encoded as stand-off markup (https://wiki.tei-c.org/index.php/Stand-off_markup). In the TEI document, we only annotate the reference to a lemma in Grassmann’s dictionary, which is also available in TEI format through an API. The motivation for modeling all data in one document is, on the one hand, to show all the data per stanza in one single source and, on the other hand, to simplify the process of exporting the data in TEI format. To enable data export for users is a key feature that is planned to be implemented in *VedaWeb* in the next project phase. In the web application, TEI data is imported and stored in a document-oriented database (mongoDB, see https://www.mongodb.com/), largely reflecting the original structure of the TEI model. This allows for fast data delivery and flexible data manipulation. Additionally, the use of a database is a prerequisite for implementing personalized work spaces for researchers to add, edit and share data, which requires at least a combination of versioning functionality (revision history) and user management.

To ensure the sustainability of the data integrated and software created within the project, *VedaWeb* collaborates with the Data Center for the Humanities (DCH), which is a “CLARIN Knowledge Centre for linguistic diversity and language documentation” (see http://ckld.uni-koeln.de/). Furthermore, the DCH is a CLARIN C-Centre (https://centres.clarin.eu/centre/47) and has applied for B-Centre status

(currently in the certification process). As such, the DCH hosts the data of the *VedaWeb* project, ensures their long-term archiving and provision, and will also take care of the data and software produced within the project beyond the end of the funding period.

4 Background and Starting Points

Before we address more specifically the challenges and opportunities arising in this collaboration, we sketch the general structures and dynamics of collaboration within the *VedaWeb* team, which play a key role for the success of the project.

The project team has been on a steep learning curve with regard to reaching a shared language and understanding of the tasks and challenges arising in the project for the different team members. This is not only true of the chasm between the "linguistic" and "technical" camp respectively, but applies on a more fine-grained level. The following diagram illustrates the fields of competence (narrow bars) as well as the de facto work areas (thick bars) of the various team members and primary collaboration partners. The diagram makes clear that the team members do not actually fall into two neatly separated domains, but that the range of competence areas are carved up in different ways with overlapping responsibilities.

Although the project has a rather small staff basis financed by external funding (DFG, German Research Foundation) with only one full researcher position (shared among two DH specialists, BK and FM), two student assistant positions (NK and JH) and a limited amount of money for the transition into a permanent hosting and curation by the data center (JB), it is supported by a large team of researchers involved either as principal investigators (UR, DK, PS, CN) or as collaborators in Cologne (FR) or abroad (PW). Further colleagues such as K. Ryan (Harvard) and D. Gunkel (Richmond) and consultants for special issues such as M. Gödel (Cologne) for the TEI data model have also contributed to the project, but are not included in the diagram for space constraints.

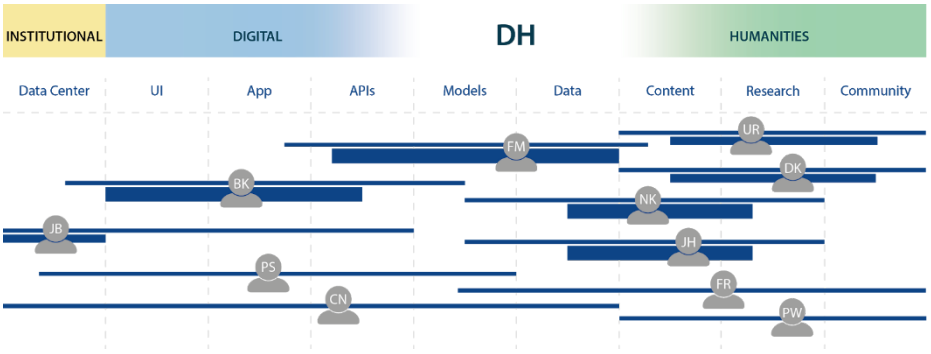

Fig. 1. Roles and coverages. Family constellation for the collaboration in the *VedaWeb* project.

As illustrated by the diagram, which we informally call our “family constellation”, there is no simple bifurcation between digital humanities and linguistics. Rather, each team member contributes to a certain range of domains, several ones of which are not clearly or not purely either technical or linguistic. This differentiation and, crucially, the overlap of competence areas among team members makes this project possible in the first place, as no single team member could be competent in or responsible for the whole range of domains.

In addition to differentiation and overlap, regular communication and close feedback loops ensure productive, agile and goal-oriented work dynamics, the success of which is mirrored by the fact that, half of the funding phase having elapsed, the project is exactly in time with its planned schedule. Most online communication, collaboration and data exchange takes place in GitLab (https://about.gitlab.com/), where several of the linguistic bodies of data are stored, too. Using GitLab's issue tracking system, team members are able to quickly provide feedback on each other’s work and effectively react to changing requirements in the project.

Team meetings in person take place on a regular basis (roughly once a month on average). In this way, developments can be steered into the right direction, mis-communications can be detected early, and there is room for creativity and adaptation. In fact, since the project is on time and running smoothly, the team is already addressing certain steps that were originally believed to be beyond the first funding period, such as the inclusion of further text versions and translations. The overlapping structure of the competence areas among the team members, paired with the regular meetings, which are supplemented by highly regular (often daily) online communication and collaboration, give rise to a sense of plasticity, i.e. to a dynamic, adaptable development of the project. Given the outlined project dynamics, the intermediate targets planned in the original application are reached so far and, at the same time, the project development can be flexibly adapted and expanded to find ideal solutions.

4.1 Simple and Challenging Issues

“What was easy and what was difficult – and why?” In general, team members at the more technical and at the more linguistic pole were frequently surprised at what exactly it was that was either easy or difficult for the others. To give some examples, implementing multiple, combinable full-text searches was easily realisable, which to a linguist who does not normally work with the support of elaborate digital search functions was surprising. At the same time, the fact that it would be technically quite difficult to have search functions run over diversely structured sets of data was not expected. Conversely, from the point of view of the DH researchers, the significant complexity of the Rigveda into different sub-structures including books, hymns and verses of different length and partially different type turned out to be a surprisingly complex and challenging data basis.

In the case of these and other examples, it was only possible to identify what was easy and what was difficult, and act accordingly, through the outlined constant and intensive communication among all project members. Furthermore, teasing out

exactly the level of complexity and time requirements for individual tasks was enabled through the structure of differentiation and overlap with regard to competence areas and responsibilities of the different team members.

4.2 Mutual Change of Views

"How did researcher and technician change each other's way of looking at things?" The linguistic researchers gained considerable new insights into the opportunities afforded by a digital research platform endowed with powerful search mechanisms and an integration of textual with lexicographic data. In developing the *VedaWeb* platform, the linguists gained a much deeper understanding of the affordances of building both a web-based application and a TEI model for the purpose of ensuring data longevity. On the other hand, the technical researchers gained new insights into the complexities of ancient texts and further linguistic resources, both with regard to their respective internal structures as well as the linguistic needs of accessing those resources. Numerous ones of these insights gained into each other's fields are likely to benefit the project members also above and beyond the *VedaWeb* project, as they are representative of DH work and historical, corpus-linguistic work respectively.

4.3 Disciplinary Blind Spots

"Did they, for instance, make each other aware of blind spots they had?" From day one of the *VedaWeb* project, the researchers and DH experts were in a constant dialogue that frequently uncovered blind spots in various domains and on all sides. Detecting these blind spots and addressing them is considered by the project team as central to the hitherto significant advances in building the *VedaWeb* platform. In particular, the repeated discussion of certain issues in much detail and in both professional and laymen's terms to include all team members in the best possible way proved essential to reaching the goals of the first project phase.

For all team members, it proved essential to clarify the technical and data-oriented terminology and the needs of both DH specialists and humanities scholars to establish data models capable of reflecting the complexity of linguistic data. For instance, the complexity and variety of the textual layers (including different versions of the original text according to different orthographic conventions and sub-structures), translations as well as levels of linguistic annotations (e.g. morpho-syntactic, lexical, metric, lexicographic) was a major blind spot for the DH specialists. One challenge in this regard is that the linguists often possess tacit knowledge of certain structural facts and it takes an extended and detailed communication process to make these explicit, or to even realise the need for explicitness. In some ways, the linguists have only gained a deepened understanding of the crucial need for explicitness for the purposes of modelling in the course of the project. There have also been various further domains in which blind spots became apparent, such as the need to address the philological genesis of the texts or the need to clarify copyright and other legal issues. In this regard, the linguists have been on a learning curve given the differences in

requirements of a web-based resource in comparison to their normal academic outlets (i.e. publication of books or journal papers).

4.4 The Whole Is More than the Sum of Its Parts

"Did the combination of thinking from a DH research question and thinking from a technical solution lead to new insights?" The integration of textual and lexicographic data, as well as various types of linguistic annotation layers into a digitized and standardized format modelled in TEI revealed manifold inconsistencies and gaps in the input. In order to address these challenges, a detailed correction of, e.g., mappings between textual data and lexical entries has been ongoing since the beginning of the *VedaWeb* project. In this way, the transfer to digital formats has led to a significant and necessary increase in quality of the linguistic data.

Besides the new possibilities with regard to, e.g., combined searches across multiple layers of text and annotations, the digitization of the Rigveda also opens up fields that have hitherto been hardly accessible to specialists of Old Indo-Aryan texts. This is true, for instance, of semantic searches that will enable a novel type of computerized research by linguists, philologists, philosophers, and other specialists of Old Indo-Aryan texts.

4.5 Towards Improved Collaboration

"How could better training or education of scholars and digital experts make collaboration easier, more effective and more efficient?" DH researchers collaborate with researchers from manifold disciplines, so that it seems unrealistic to include in their training specifics of any one humanities discipline. However, DH specialists need a general understanding of the objects of study in the humanities, of research questions and research methods, knowledge of which should certainly be included in their training. As the digital experts are all graduated in Digital Humanities (instead of, e.g., computer science), the project team members of *VedaWeb* have such training and in part previous experience with specifically linguistic projects, and were accordingly ideally prepared. Any specific insights that go beyond this general background was then gained in the course of the project through the outlined processes of regular and detailed communication and collaboration.

Similarly, detailed training in DH methods, approaches and resources may not be possible or necessary to include in a linguistics training. However, a general understanding of the different fields of DH work is certainly highly desirable, especially given the intense, ongoing digitization of research in the humanities disciplines. For instance, the fact that there is a significant difference between building a web application, on the one hand, in contrast to ensuring sustainability of data by modelling them in TEI, on the other hand, is an insight into the nature of DH projects that is likely to benefit collaboration from early project stages. The insufficient insights of the humanities researchers into these components of technical DH work was a hindering factor in the early stages of *VedaWeb* and a training for humanities researchers that includes a general understanding of these and other

general structures and processes is likely to be beneficial for collaborative work across disciplines. In particular, this would involve developing a deepened understanding for the need of modelling and formalizing knowledge structures, making them explicit and thus usable for computational implementation. Moreover, there is the need for a basic understanding of the interaction of different technical components including data formats, databases, search engines, program logic, web interfaces and usability.

5 Conclusion

We have described an instance of the "humanities problem" for linguists of handling large textual corpora, in this case the corpus of Old Indo-Aryan texts. In this project, the "technical solution" consists of developing a formal data model for the given knowledge domain and the construction of a web-based research platform that integrates multiple layers of textual data (different versions of the original text and various translations) with linguistic annotations (e.g. morpho-syntactic and metrical annotations) as well as with lexicographic data via APIs to specialized dictionaries. Half of the funding time having elapsed, it has become clear that it is not only helpful to collaborate closely among the different team members, but that a continuous and intense process of communication, exchange and replanning is crucial for the realization of the project. In particular, regular discussion to clarify terminology, on the one hand, and the uncovering of blind spots among team members, on the other hand, have proved indispensable. By engaging in such close collaboration, the project has been able to stay in time with its planned schedule despite a long initial phase of finding a common understanding and language, which is in part ongoing. Above and beyond the communicative practices adopted, the project benefits significantly from a diverse team with a variety of backgrounds and competences. A further advantage is that the DH specialists in the team are already experienced in handling linguistic data. However, we believe that interdisciplinary project teams in general can reach a comparable level of mutual understanding when embracing particularly close collaboration [1, 5]. This involves the continuous discussion and clarification of terminology, of data formats and of models, as well as short feedback loops to create a shared understanding of humanities research questions and to reach tailored DH solutions.

References

1. Deegan, M., McCarty W. (eds.): Collaborative Research in the Digital Humanities. Routledge, Farnham (2011).
2. Geldner, K.F.: Der Rig-Veda. Aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen von Karl Friedrich Geldner. Harvard University Press, Cambridge (Mass.) (2003)[1951–57].
3. Grassmann, H.: Wörterbuch zum Rig-veda. O. Harrassowitz, Wiesbaden (1873).
4. Grassmann, H.: Rig-veda. Übersetzt und mit kritischen und erläuternden Anmerkungen versehen von Hermann Grassmann. F.A. Brockhaus, Leipzig (1876).

5. Griffin, G., Hayler, M.S: Collaboration in Digital Humanities Research – Persisting Silences. Digital Humanities Quarterly 12(1) (2018). http://www.digitalhumanities.org/dhq/vol/12/1/000351/000351.html
6. Griffith, R.T.H.: The Hymns of the Rigveda. Lazarus, Benares (1896).
7. Haig, G., Schnell, S.: Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Manual Version 7.0 (2014), https://www.academia.edu/10328418/Haig_Geoff_and_Stefan_Schnell._2014._Annotations_using_GRAID_Grammatical_Relations_and_Animacy_in_Discourse_._Manual_Version_7.0, last accessed 2019/02/11.
8. Hellwig, O., Hettrich H., Modi A., Pinkal M.: Multi-layer annotation of the R̥gveda. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7.-12.5. 2018. European Language Resources Association (ELRA), Paris (2018).
9. Hintze, A.: „Lohn“ im Indoiranischen: eine semantische Studie des Rigveda und Avesta. Beiträge zur Iranistik 20. Reichert, Wiesbaden (2000).
10. Kazzazi, K.: „Mann“ und „Frau“ im R̥gveda: mit einem Exkurs über Wörter für „Frau“ im Atharvaveda. Innsbrucker Beiträge zur Sprachwissenschaft 98. Inst. für Sprachwissenschaft, Innsbruck (2001).
11. Keydana, G., Luraghi, S.: Definite referential null objects in Vedic Sanskrit and Ancient Greek. Acta Linguistica Hafniensia 44(2), 116–28 (2012). doi: 10.1080/03740463.2013.776245
12. Kölligan, D.: Vedisch *nigut-* ‘Feind, Übeltäter’. Historische Sprachforschung 120, 134–141 (2007). www.jstor.org/stable/40849294
13. Kulikov, L.: The Vedic -ya-Presents: Passives and Intransitivity in Old Indo-Aryan. Rodopi, Amsterdam (2012).
14. Oldenberg, H.: R̥gveda. Textkritische und exegetische Noten. Weidmann, Berlin (1909/1912).
15. Reinöhl, U.: Grammaticalization and the Rise of Configurationality in Indo-Aryan. Oxford University Press, Oxford (2016).
16. Renou, L.: Études védiques et paninéennes. Boccard, Paris (1956–1969).
17. Roesler, U.: Licht und Leuchten im Rgveda. Untersuchungen zum Wortfeld des Leuchtens und zur Bedeutung des Lichts. Indica et Tibetica Verlag, Swisttal-Odendorf (1997).
18. Ryan, K.: Onsets contribute to syllable weight: statistical evidence from stress and meter. Language 90(2), 309–341 (2014).
19. Scarlata, S., Widmer P.: Vedische exozentrische Komposita mit drei Relationen. Indo-Iranian Journal 58(1), 26–47 (2015).
20. Schmid, W.P.: Die Wurzel *VI* im Rgveda. In: Mélanges d’indianisme. à la mémoire de Louis Renou, pp. 613–624. de Boccard, Paris (1968).
21. Widmer, P.: Der altindische vrkī́-Typus und hethitisch nakki-: Der indogermanische Instrumental zwischen Syntax und Morphologie. Die Sprache, Zeitschrift für Sprachwissenschaft 45(1–2), 190–208 (2005).
22. Witt, A., Rehm, G., Hinrichs, E., Lehmberg, T., Stegmann, J.: SusTEInability of linguisitc resources through feature structures. Literary and Linguistic Computing 24(3), 363–372 (2009).

I Want it All, I Want it Now. Literature researcher meets programmer

Vanessa Hannesschläger[1[0000-0003-0938-0890]] and Peter Andorfer[1[0000-0002-9575-9372]]

[1] Austrian Academy of Sciences, Sonnenfelsgasse 19, 1010 Vienna, Austria
vanessa.hannesschlaeger@oeaw.ac.at
peter.andorfer@oeaw.ac.at

Abstract. This paper describes a collaborative project carried out in 2017. The initial motivation to do the project was a call for participation in a conference on text genetic editing in digital editions. A literature researcher (Vanessa) asked a programmer (Peter) to work with her on a little publication platform which would display an edition focusing on the text genesis of a specific play written by her main subject of study, the Austrian writer Peter Handke, to present at the aforementioned conference - he agreed. In turn, Peter asked Vanessa to come up with a research question about the text and a list of features she would need the platform to have. Her answer was simple, really - and it became the title of this paper, which speaks of what Vanessa wanted, what Peter wanted, how they did it and how that worked out. We will describe the data modelling, the automated and the manual processing of the data, the tools used, the technical implementation, the resulting *handke-app* and the challenges and benefits of two very different research perspectives in all these steps.

Keywords: Peter Handke, TEI, Digital Editing, Text Genesis, Automated Collation

1 Preface

1.1 A Subsection Sample

This paper describes a collaborative project carried out in 2017. The initial motivation to do the project was a call for participation in a conference on text genetic editing in digital editions. A literature researcher (Vanessa) asked a programmer (Peter) to work with her on a little publication platform which would display an edition focussing on the text genesis of a specific play written by her main subject of study, the Austrian writer Peter Handke, to present at the aforementioned conference - he agreed. In turn, Peter asked Vanessa to come up with a research question about the text and a list of features she would need the platform to have. Her answer was simple, really - and it became the title of this paper. As its reader, you first need to understand that to a Handke researcher, everything matters: every text witness, every correction, every date, every person involved, every pen used, every coffee stain, every archive holding a Handke collection, every place in which a part of the writing process took place, every location of the play, every book read by the author, every language used in the text, every biography of every person that might have been inspiration for one of the characters of the

play, every production of the play, every person involved in every production of the play, every published version of the text, every translation of the text, and of course all other texts that Handke wrote and the cross references with this one text - to just name a few of the relevant aspects. As Vanessa had written her diploma thesis about the play in question, she already had all these data, though not in a machine readable form. When Peter asked Vanessa to focus and pick the most important aspects so that he could start developing data models and technical solutions for them, she repeated her answer. In this paper, we describe what things we managed to put into practice and where the realities of time management stopped us from constructing the Swiss army knife Vanessa had originally envisioned (and Peter had at no point in time intended to build).

2 Introduction and Scope

Peter Handke's play *Immer noch Sturm* (2010) (*Storm still*, 2014)[1] is a perfect source text for a study on modelling the genesis of a literary text: Five text stages (with several sub-stages) can be identified even before page proofs.[2] The numerous smaller and bigger adjustments Handke made from stage to stage, but also the large quantity of accompanying material (such as the author's preparatory reading notes) make this corpus a good starting point for a study on the possibilities and boundaries of sustainable encoding of text genetic processes.

In our study, we transcribed the respective first page of each text witness and added TEI P5 markup. The result of our work was published under the title *handke-app*.[3] The encoding focused on issues such as the distinction between immediate and later manual corrections or the representation of the proven integration of preparatory reading notes and thus served as a list of requirements for functionalities of the web application which would represent these encoded texts. The goal of this application is to provide the best possible support for the analysis and research process.

3 Material

Storm still is a formally complex text for the stage about the Carinthian Slovene resistance against the Nazi occupation during the Second World War. The *dramatis personae* include a nameless "I", "who is not reasonably distinguishable from the author's persona"[4], and his Carinthian Slovene maternal relatives. Its formal structure is remarkable and shows the author's continuous development of his own approach to "epic theatre", in which he combines elements of the antique tragedy with epic narrative from a first person perspective.[5]

The text genetic material is as remarkable as the text itself. The first version was written from December 15th, 2008 to February 22nd, 2009 (in pencil). Before the page proofs were created in July 2010, four further text stages with several sub-stages emerged due to continuous adaptations and changes to the text made by the author. In addition to this extensive (and fully preserved) material, additional notes, books, and other materials (all kept by the Salzburg literary archives) give further insights into the process of creation.[6]

The situation described above poses a number of questions and problems. These lead to the question that inspired our project: What does a digital edition have to be able to do in order to serve a literature researcher's needs? The answer: It has to represent all existing knowledge and research about the edited text, thereby generating new knowledge.

However, "all" existing knowledge is a relative term in this context. We could have focussed on the work context - the play and its relations to Handke's other works -, the biographical context - the characters and their relations to Handke's real life family members -, the historical context - the representation of historical events and the text's relations to sources about these events[7] -, or the text genetic context. The latter is a good choice for various reasons, one of them being the availability of information and data about the text genesis via the platform Handke*online*.

4 Encoding

Even though the platform Handke*online* is a data treasure for Handke research, it has some clear disadvantages from a technical perspective. It does not provide the possibility to extract any structured data, let alone process it. Therefore, quite some manual work was necessary.

The following data were relatively easily transformable into structured data:

- Dates: Handke has been documenting the writing dates of his first text versions for many years, and this is also true for *Storm still*. Every writing date is noted in the manuscript next to the text written on the respective day. Usually (and also in this case), he also notes the dates on which he worked on subsequent versions of a text.
- Persons and institutions: Information on people and institutions formally involved in the production of a text stage (e.g. transcriber, editor, owner of the manuscript) had already been collected by Handke*online* and could therefore be transformed into structured data easily.
- Places: Thanks to meticulous documentation of writing places in the manuscripts, the identification and subsequent geo referencing of places relevant to the genesis of the text was unproblematic.

On the other hand, the translation of the following information into machine-readable form posed certain challenges:

- Preparatory reading: In preparation for writing *Strom still*, Handke read extensively about the history of the Carinthian Slovene resistance against the Nazi occupation of Carinthia during World War II. His reading focused on partisan memoirs. The most important book read in this context was Karel Prušnik-Gašper's *Gemsen auf der Lawine* [Chamoix on the avalanche].[8] Handke's triple reading of this book (each several years apart) can be dated exactly as the book, in which he made notes and annotations in different colors during each reading and also noted beginning and end dates, is available in the archives.[9] For other books, the data is more fuzzy: He took notes and collected quotes, which are partly dated and have been preserved, however the books themselves have not. In addition, further readings of books to which no notes have been preserved can be proven by identification of

direct quotes - however, the time of his reading can only be estimated. This is one of several examples of substantial information about the text genesis which has been investigated and confirmed by research, but can still not be pinned down and transformed into precise, machine-readable data.

- Source indication: The research articles about Handke's *Storm still* quoted above show that the reconstruction of the text's genesis was a task of many years. Vanessa, one of the authors of this paper, has worked on this text since it was published and dedicated her diploma thesis as well as several papers and a lot of her work done for Handke*online* to the investigation of its becoming, its meaning, and its interconnections to other texts. She was also the data provider for the app we developed. Thus, we did not source our data from other research, but rather deduced it from the data provider's previously accumulated knowledge about the topic. For this reason, we did not add a file including bibliographic information on sources used for the app. Another reason was the additional time it would have required to connect such a file to the individual data points. While we are confident that this was a legitimate decision in terms of research ethics, we still see the problem that the app does not indicate if a given data point was sourced from Handke*online*, Vanessa's thesis or a paper.
- Provenance history: Even though all text genetic material for *Storm still* is kept by the Salzburg literary archives today,[10] it was previously owned by several individuals and therefore arrived at the archives not as one collection, but in parts and with time. In addition to this, only parts of the material belong to the archives, other parts are privately owned and only kept by the archives as a permanent loan. The history of this material and its paths is complex, but it is known - in principle. This is informal knowledge among Handke researchers and fans. Therefore, reliable and exact data suitable for structured analysis cannot be deduced from this information which some might refer to as gossip - it might be that from a present perspective, but looking, say 200 years to the future, this might be valuable information for researchers who could be interested in the author's network or the market prices for manuscripts at the beginning of the millennium. In the long run, not preserving this information might therefore mean a loss.

The transcriptions of the various text stages are previously nonexistent data which were newly created for the *handke-app*. As our goal was not to provide a full edition of *Storm still*,[11] but rather a technology test, we only transcribed the respective first pages of each text witness and encoded it with TEI-P5 markup[12].

5 General Set-up

The described project was a pilot study. Work on the research content of the project, i.e. development of the research question, transcription, and annotation (edition) of the text witnesses, was carried out by Vanessa. Peter was responsible for all technical aspects as well as development of the data management workflow. Both contributors are employees of the *Austrian Centre for Digital Humanities* of the *Austrian Academy of Sciences* (ACDH-OeAW) which provided the necessary (server) infrastructure.

6 Document Centered Approach

The outcome of the project was shaped by the team's decision to follow a document centred approach for the transcription. Each text witness was transcribed in an individual XML file. The <teiHeader> elements of each file contain the metadata specific to the text witness (archive holding the manuscript, physical traits, history of its genesis) structured in a TEI conformant way (as far as that was possible). The next step was to transcribe the respective first page of the text witness and encode specific text genetic phenomena within the respective text witness and model the formal structure of the text (using <pb>, <lb>, and <p>).

As the pilot study focused on text genesis, we refrained from encoding text genre specific phenomena according to the TEI module *Performance Texts*[13], e.g. indication of a <speaker>, or a more in-depth literature analytic markup e.g. providing information on time, place, or characters, as we found this to be insignificant due to the incompleteness of the transcription.

We encoded the work's genesis, i.e. the systematically documented changes, deviations, and variants, in the next step, which was collation. While methods of collation strongly vary between disciplines, research projects, and even individual researchers in "analogue" text research[14], it is a strongly formalized approach in the digital humanities. This is necessary as collation is in this case primarily carried out by machines. According to the *Gothenburg model*[15] developed in 2009, the following steps have to be taken:

1. The respective witnesses have to be divided into comparable chunks of text; this process is called *tokenization*. This is generally done on the word level (which the machine defines by identifying strings of symbols separated by spaces).
2. In a second step, the tokens of the respective witnesses are compared to each other. For the (likely) case of differing amounts of tokens, so-called *gap tokens* have to be inserted.
3. Based on this comparison, analysis can be carried out. However, the authors of the *Gothenburg model* have pointed out that this task might be beyond the machines' limits, especially when it comes to the task of identifying how deviations in various witnesses are related to each other and if differing sequences of tokens are additions, deletions, or transpositions of text parts: "While alignment results can still be judged in terms of their quality to some extent, transposition detection can only be done heuristically as one can easily think of cases, where it is impossible for a computer 'to get it right'."[16]
4. The final step is the synthesis if the results of collation. Just as in "traditional" text studies, the result can be a critical apparatus that documents deviations from a base text version in other witnesses. Depending on the technical implementation of the *Gothenburg model*, it can also be a graph and/or a tabular representation.

The project team was aware of two concrete implementations of this model, namely *CollateX*[17] and *Juxta Commons*[18]. The decision for *Juxta Commons* was made due to more user friendliness (which helped Vanessa in doing her part), i.e. *Juxta Commons*

is a web service with a graphic user interface while *CollateX* requires a local installation and some familiarity with use of the command line.

With the help of *Juxta Commons*, Vanessa was able to collate the individual text witnesses and thus to encode (or let encode) the text genesis in a quick, systematic, and machine-readable form. The genesis of the text was annotated by *Juxta Commons* according to the *Parallel Segmentation Method*[19], which is characterized by the notation of the various readings next to each other (parallel), which facilitates the comparison of variants. A short example:

```
<app>
  <rdg wit="#wit-24801 #wit-24800">eine Sitz-</rdg>
  <rdg wit="#wit-24794">Eine Sitzbank</rdg>
  <rdg wit="#wit-24795">Eine<lb/> Sitzbank</rdg>
  <rdg wit="#wit-24793">Eine Sitzbank</rdg>
  <rdg wit="#wit-24798">Eine Sitzbank</rdg>
  <rdg wit="#wit-24799">Nichts</rdg>
  <rdg wit="#wit-24796">Eine Sitzbank</rdg>
  <rdg wit="#wit-24797">Eine Sitzbank</rdg>
</app>
```

These results were manually cleaned in order to obtain better readability by wo/man and machine:

```
<app>
  <rdg wit="#wit-24801 #wit-24800">eine Sitz-</rdg>
  <rdg wit="#wit-24794 #wit-24793 #wit-24798 #wit-24796
    #wit-24797">Eine Sitzbank</rdg>
  <rdg wit="#wit-24795">Eine<lb/> Sitzbank</rdg>
  <rdg wit="#wit-24799">Nichts</rdg>
</app>
```

Here we encountered one of the challenges of communication and coordination between a “tekkie” and a “human”: This code optimization could easily have been done via a small script instead of manual cleaning, had Vanessa thought to ask for it. However in addition to this, Vanessa also had to manually add in manuscript corrections by the author which had been encoded previously, but which *Juxta Commons* failed to include in the collation. An example of this is the <subst> element in the following passage:

```
<app>
  <rdg wit="#wit-24799">I</rdg>
  <rdg wit="#wit-24794 #wit-24795 #wit-24793 #wit-24801
#wit-24800">EINS</rdg>
    <rdg wit="#wit-24798 #wit-24797">ERSTER AKT</rdg>
    <rdg wit="#wit-24796">
      <subst>
        <del>ERSTER AKT</del>
        <add hand="#handke" place="below">EINS</add>
      </subst>
    </rdg>
```

</app>

Looking back, it would have been more efficient to do without text critical markup in the first transcription and only adding it in after collation; but as the team had never worked with this tool before, we were not aware of this problem beforehand.

The reason for the choice of this text witness centred approach was ultimately of a technical / pragmatic nature: For encoding, we used the *oXygen*[20] XML editor for the very simple reason that Vanessa was already familiar with using this tool and the ACDH-OeAW owns licenses for it.

7 Implementation

The web application for the *handke-app* was implemented using *exist-db* for the following reasons: As the transcriptions and annotations were done in XML format, use of a native XML database stood to reason. Additionally, *exist-db* can easily be integrated in *oXygen*; and last, but not least Peter is experienced in working with *exist-db* and its functionalities that facilitate the development of data-driven web applications.[21]

The following features were successfully implemented in the *handke-app*:

- Text views 1[22]: Access to individual text witnesses via a traditional table of contents. Individual text views include extensive meta information (document title, archive holding the manuscript, original title, transcriber, license) as well as the text (including additions, deletions, etc.). A scan of the original manuscript page is also available. [23]
- Text views 2[24]: This entry point allows users to see text witnesses next to each other in order to compare them. For this view, the *EVT Viewer*[25] was used. From a technical perspective, it was very pleasant to see that the *EVT Viewer* was able to process the files created by Juxta Commons ad edited by Vanessa without further adaptations of the application's code.
- Text views 3[26]: The result of the *Juxta Commons* collation can also be exported as a "traditional" apparatus in a static HTML file.
- Indices 1[27]: Events. This page collects and lists meta information retrieved from the <teiHeader>s of all XML files. Thus, all events related to the text genesis (individual writing days, Handke's preparatory reading of books, corrections by Handke and others, etc.) are collected in a list here. In addition, they were visualized on a map with an included timeline. Due to the mentioned inaccuracies and problems of standardizations of certain informations (e.g. precise dates of preparatory reading), this visualisation's meaningfulness is limited.
- Indices 2[28]: Persons. This page lists all persons involved with the material, be it as owner or in the creation process (transcriber, editor, etc.). Persons are attached to a location (where their contact with the text took place). These locations are visualized on a map.
- Indices 3[29]: Places. The same map as on the person page and a list of the places.

- Indices 4[30]: Institutions. A list of all institutions involved. Indices 3 and 4 are not particularly useful due to the small amounts of data, but were so easily implementable from a technical point of view that we decided to include them anyway.
- Analyses 1: Deletions and additions. Two graphs show the amount of deletions and additions (i.e. the frequency of the TEI tags <del> and <add>) in all text versions.
- Analyses 2[31]: User requests. Users can choose a TEI tag and query for its frequency in all text versions. As only a small amount of tags was used in the project at hand, this feature is not particularly useful for this specific data set. It was included because Peter wanted to test the necessary efforts of implementing this feature.

8 Conclusion

Summing up, we can conclude that this pilot study was fruitful both for the "tekkie" and for the "human". While there were challenges in communication and cooperation at certain stages, we managed to broaden each other's view and understanding considerably and both benefited from the cooperation.

Our work showed that a complex work genesis including a number of text witnesses can be encoded efficiently by following a text witness centered approach and subsequently using machine supported collation. This is especially true when the result can be processed using existing software such as the *EVT Viewer*.

Another positive result is that we were able to show that digital methods can provide modes of analyses (i.e. quantitative queries about text phenomena) that a traditional apparatus would not be able to offer (even though the result was not meaningful in the case of our pilot study due to the limited amount of data).

We also learned that the *Text* Encoding Initiative's guidelines, while the unquestionably best approach for encoding text inherent phenomena, reach their limits when used for encoding "real world phenomena" related to text genesis such as places, persons, or events. Even though the TEI offers elements for encoding these phenomena,[32] the interconnection of these entities to each other and to the text witnesses is not well specified and largely varies from project to project. Therefore, connecting the data created in the pilot study to other projects will likely be difficult to impossible. Use of a more comprehensive model that does not only focus on encoding text, but also extra textual realities (e.g. *CIDOC CRM*[33]) might have been a better choice.

From the literature researcher's point of view, the pilot study was also very fruitful. Though the result is not the "Swiss army knife" originally imagined, it is a pretty nifty tool that can do some unanticipated things. While sometimes challenging, the methods and tools used were manageable even for a non-"tekkie" and worth every effort considering the results. The visualizations and features inspired a deeper understanding of the text and its becoming as well as completely new perspectives - even though Vanessa thought that she knew this text inside out even before she started doing this study, hav-

ing worked on it for numerous years. We will spare you the explanation why it is amazing to learn that Handke only transformed the "ninety nine apples" on the apple tree mentioned in the first stage direction of this play into "99 apples" in the very last text witness. But believe us: it truly is mind blowing, and we would never have found this out without the *EVT Viewer*.

From the developer's point of view, the pilot study was fruitful in regards of getting a deeper understanding of the "Swiss army knife" metaphor. On a first glimpse, a Swiss army knife is one single tool which can do many things. Though on a second look it could also be understood as a collection of many different kind of tools wrapped up between to red plastic halfs. Transposing this metaphor into the digital (humanities) world, we reach the idea of many (micro) services/tools, each one tailor made for a single task (i.e. the *oXygen* XML editor for encoding text, *eXist-db* for storing data, *Juxta Commons* for collating text) "glued together" by some basic website, providing links to all those services / tools / data. The main challenge in the interaction with the non-"tekkies" is the communication process needed to break down a huge research question like "I want it all" into its many components. This is doable. But needs time and patience.

Finally, we have to mention that since this pilot study, we have worked together on various projects and gotten better at understanding each other's perspectives and languages better and better over time. Our cooperation is ongoing, as is the work on the *handke-app*, as we noted on its start page: "Please be aware: This is work in progress. If you find any mistakes or have suggestions for further development, please create an issue in the project's code-repo on GitHub."[34]

References

1. Handke, P.: Immer noch Sturm. Suhrkamp, Berlin (2010), respectively Handke, P.: Storm Still. Trans. Chalmers, M. Seagull, London, New York, Calcutta (2014).
2. See the list of text genetic material for *Storm still* on Handke*online*, http://handkeonline.onb.ac.at/node/57/material, last accessed 2019-02-04.
3. handke-app, https://handke-app.acdh.oeaw.ac.at/, last accessed 2019-02-04, respectively acdh-oeaw/handke-app: First release (Version v1.0), http://doi.org/10.5281/zenodo.1195978, last accessed 2019-02-04.
4. Kastberger, K.: Lesen und Schreiben. In: Kastberger, K., Pektor, K. (eds.): Die Arbeit des Zuschauers. Peter Handke und das Theater. Jung und Jung, Vienna, Salzburg (2012), pp. 35–47, p. 44 [transl. VH].
5. See Hannesschläger, V.: Real Life Fiction, Historical Form: Peter Handke's "Storm Still". In: Boldrini, L., Novak, J. (eds.): Experiments in Life-Writing. Intersections of Auto/Biography and Fiction. Palgrave, London (2017) [Palgrave studies in Life-writing 1], pp. 145–165.
6. See Hannesschläger, V.: „Geschichte: der Teufel in uns, in mir, in dir, in uns allen." – Zur Rezeption von Familiengeschichte und Historie in Peter Handkes Immer noch Sturm. Vienna (2013) [Diploma thesis].
7. See Hannesschläger, V.: Peter Handkes „Immer noch Sturm" und Karel Prušnik-Gašpers „Gämsen auf der Lawine". In: Wieser, L. (ed.): Karel Prušnik-Gašper: Gämsen auf der Lawine. Materialien. Wieser, Celovec (2016), 13–18.

8. Prušnik-Gašper, K.: Gemsen auf der Lawine. Der Kärntner Partisanenkampf. Drava, Celovec (1980). Handke read and annotated this edition himself (see Handke*online*, http://handkeonline.onb.ac.at/node/1566, last accessed 2019-02-04).
9. Handke*online*, http://handkeonline.onb.ac.at/node/1566, last accessed 2019-02-04.
10. Salzburg literary archives, collection: Handke, Peter (LAS) and collection: Handke, Peter (Leihgabe Widrich) (PH-PAW).
11. This was infeasible due to copyright reasons on the one hand, and limited (time and financial) resources on the other.
12. TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange (2017), http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf, last accessed 2019-02-04, respectively http://www.tei-c.org/release/doc/tei-p5-doc/en/html/, last accessed 2019-02-04.
13. See TEI Consortium: TEI P5: Performance Texts, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html, last accessed 2019-02-04.
14. See e.g. Sahle, P.: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe. Norderstedt (2013) [Schriften des Instituts für Dokumentologie und Editorik – Band 7].
15. TEI SIG Manuscripts: The "Gothenburg model": A modular architecture for computer-aided collation (2011), https://wiki.tei-c.org/index.php/Textual_Variance, last accessed 2019-02-04.
16. Ibid.
17. CollateX, https://collatex.net, last accessed 2019-02-04.
18. Juxta Commons, http://juxtacommons.org, last accessed 2019-02-04.
19. TEI Consortium: TEI P5: Parallel Segmentation Method, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS, last accessed 2019-02-04.
20. SyncroSoft: oXygen XML editor, https://www.oxygenxml.com/, last accessed 2019-02-04.
21. Andorfer, P.: dsebaseapp (2016ff), https://github.com/KONDE-AT/dsebaseapp, last accessed 2019-02-04.
22. https://handke-app.acdh.oeaw.ac.at/pages/toc.html, last accessed 2019-02-04.
23. The scans were kindly provided by the Salzburg literary archives.
24. https://evt.acdh.oeaw.ac.at/#/critical?d=doc_1&e=critical, last accessed 2019-02-04.
25. Edition Visualisation Technology – EVT Viewer, http://evt.labcd.unipi.it/, last accessed 2019-02-04.
26. https://handke-app.acdh.oeaw.ac.at/pages/juxta-play.html, last accessed 2019-02-04.
27. https://handke-app.acdh.oeaw.ac.at/pages/events.html, last accessed 2019-02-04.
28. https://handke-app.acdh.oeaw.ac.at/pages/persons.html, last accessed 2019-02-04.
29. https://handke-app.acdh.oeaw.ac.at/pages/places.html, last accessed 2019-02-04.
30. https://handke-app.acdh.oeaw.ac.at/pages/organisations.html, last accessed 2019-02-04.
31. https://handke-app.acdh.oeaw.ac.at/pages/stats-dynamic.html, last accessed 2019-02-04.
32. See TEI Consortium: TEI P5: Names, Dates, People, and Places, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html, last accessed 2019-02-04.
33. CIDOC Conceptual Reference Model (CRM), http://www.cidoc-crm.org/, last accessed 2019-02-04.
34. https://handke-app.acdh.oeaw.ac.at/pages/index.html, last accessed 2019-02-04.

Interdisciplinary Collaboration in Studying Newspaper Materiality

Eetu Mäkelä[1][0000−0002−8366−8414], Mikko Tolonen[1][0000−0003−2892−8911], Jani Marjanen[1][0000−0002−3085−4862], Antti Kanner[1][0000−0002−0782−1923], Ville Vaara[1][0000−0001−7924−4355], and Leo Lahti[2][0000−0001−5537−637X]

[1] Department of Digital Humanities
University of Helsinki, Finland
first.last@helsinki.fi
[2] Department of Mathematics and Statistics
University of Turku, Finland
first.last@utu.fi

Abstract. This paper presents a collaboration between computer scientists, linguists and historians studying the material aspects of newspapers and developing a tool for that purpose. The paper describes how the back-and-forth collaboration in terms of research questions and technical challenges yielded insights both for solving computational problems as well as refining historical analysis. In the project, existing metadata was amended by reconstructing new materiality data from the Finnish digitised newspaper corpora. The analysis of such data is crucial for studying the development of newspapers, but can also inform other computational studies on the same data. The use of enriched materiality data allows for better understanding subdivisions in large corpora such as digitised newspapers, but also highlight that content and form interact. Content analysis of newspapers should therefore always take into account material properties of the studied material to properly grasp the cultural, social and political meanings embedded in the sources.

Keywords: Materiality of newspapers · Collaboration · Digital humanities.

1 Introduction

This paper offers a view to the collaboration undertaken at the Helsinki Computational History Group (COMHIS)[3] between computer scientists, historians and linguists on a project that studies the material dimensions of newspapers and their development [3].

The present day transformation from print to digital is not the first time newspapers have evolved drastically. Instead, this change of format reminds of similar transformations when the newspaper first appeared as a distinct material genre. One influential definition separating a newspaper from a newsbook or

[3] http://helsinki.fi/computational-history

pamphlet in its early days was that a newspaper was a "sheet of two or four pages, made up in two or more columns" [10]. The Dutch had two-column news at the time, while civil war in Britain saw both the rebels and the crown printing their propaganda. It took, nevertheless, centuries before journalism became a profession of its own and newspapers took their particular shape in the mid-nineteenth century [20,1,2,11,13,23].

In the context of digital humanities, newspapers have become an iconic example of "big data" research (cf. [5,15,7], https://numapresse.org/). While in localised research [8,28] the material can be thought uniform, in the big data approaches it is striking how little attention is paid to what the data consists of. A telling example of waking up to this is the Oceanic Exchanges project (https://osf.io/wa94s/) where M.H. Beals and Ryan Cordell quickly concluded that mapping metadata across its many datasets is to be one of its most important contributions (https://twitter.com/ryancordell/status/1001845719341285377).

Framed against this background, the idea of this paper is to outline how we developed a tool to uncover and explore the varied materiality of newspapers. As part of the large-scale digitisation, the accessibility of historical newspapers has improved drastically, but at the same time much of the information about the size, shape and feel of the newspapers, that was so central to past readers in understanding what kind of documents they were perusing, has to a large extent been hidden from view. Interestingly, the digitised versions of the newspapers also allow for large-scale study of their material dimensions – an opportunity that has so far been paid very little attention to. In our case, our focus on materiality is also just one aspect of the group's larger interest in studying the nature of early modern public discourse through the analysis of structured and unstructured data relating to newspapers and other printed materials.

In what follows, we will first briefly explain the background for this study and how it fits the group's publication history. Then, we'll shortly discuss the type of data we started our work from, before going into detail on how the research process that led to the materiality explorer tool actually happened. Finally, we will describe the tool itself and the tentative results we've obtained using it, before concluding by outlining directions for future work.

2 Studying the Materiality of Newspapers

The first time that data on the materiality of newspapers was extracted and studied by us at the COMHIS group was as part of the Helsinki Digital Humanities Hackathon of 2015[4]. After that, intermittent analyses on both the content as well as metadata such as language, location and form of the newspapers was done as part of the internal dialogue of the research group, in part in the context of the Academy of Finland funded project on "Computational History and the Transformation of Public Discourse in Finland, 1640-1910"[5].

[4] http://heldig.fi/dhh15

[5] http://www.aka.fi/globalassets/32akatemiaohjelmat/digihum/hanke-esitteet/salmi-digihum.pdf

Slowly, these explorations coalesced into multiple conference presentations on the subject. Mostly, the actual work happened in sporadic bursts, often with one of the more computationally oriented researchers in the group being inspired to run a particular analysis, which then led to back-and-forth exchange between the historians and the experts in quantitative methods to better interpret and fine-tune the analysis. In this process analyses were also designed to be more aligned with research questions pertinent to newspaper history, and new analyses were requested by the historians.

In time, these explorations led to more focused research questions, dealing with the modernisation of newspapers in Finland in two main languages. As newspapers became more frequent, more topical and gained a larger format, they started resembling the modern newspapers that we encounter today (or perhaps those of our childhood). In particular, we wanted to trace the asynchronicity that was present between Finnish-language and Swedish-language papers. Editors and other intellectuals in Finland operated mostly in both languages, and thus the newspapers were developed in constant cross-fertilisation across the language border, but still the different language spheres developed at different paces. While Swedish-language papers were generally more advanced up to the 1860s and 1870s, Finnish-language papers became leading by the turn of the century 1900 due to growth both in terms of readership and places of publication.

A problem with our early explorations was that they had been done in a haphazard, off-the-cuff manner by different people using different versions of the data, so they were not mutually consistent and reliable. An impetus to change this came when one of the conference presentations led to an invitation to write up the work more formally for the Journal of European Periodical Studies (JEPS). At this point, it was decided to take one single version of the data as the source, and calculate all material and linguistic indicators from that. A more thorough analysis of the trustworthiness of the pipeline and the dataset itself was also undertaken.

For the JEPS article, the figures and analyses used to inform the content started as those that had arisen organically as part of the internal dialogue within the group. However, when polishing the art, a dialogue was held between the historians and the statistical visualisation experts on what the core message was. This led to replacing earlier more explorative versions of the visualisations with ones designed specifically to convey particular arguments. At the same time, the visual outlook of all graphs was unified.

After working on the JEPS article, the group had a relatively good notion on what the important aspects of materiality in the data were, and how they could best be visualised and explored in a unified manner. This led way to the development of an interactive materiality explorer. Through this, there was more freedom for the content experts to explore the phenomenon, with much less frequent need for the computer scientists to run customised analyses or change the parameters of the exploration.

2.1 Extracting and Deriving Material Aspects from ALTO XML

In order to understand what the group was working with, it is relevant to understand the usefulness of ALTO (Analyzed Layout and Text Object, https://www.loc.gov/standards/alto/) files that were luckily available for the project. ALTO files contain a description of the visual organisation of content on a page, at the core of which are the individual words and their page coordinates. At the same time, the words are also grouped into blocks, often corresponding to paragraphs or columns. The format also contains general layout information, such as the sizes of margins and main printed area.

The usefulness of ALTO for analysing materiality crucially depends on the choice of the measurement unit in which all coordinate and size information is given. Here, the format gives a choice from three options: mm10 (tenth of a millimeter, the default value), inch1200 (1200th of an inch) or pixel. Of these, the first two directly relate all measurements to actual physical dimensions, while the pixel coordinates do not. However, even then, the information on original physical dimensions can be recovered if the DPI value of the image is known, information given in the METS metadata files originally often accompanying the ALTOs. Unfortunately, many collections such as the Dutch Delpher (https://delpher.nl/) and French Gallica (https://gallica.bnf.fr/) provide their ALTO data specifically using pixel coordinates, while not giving out the METS files (which would also contain logical segmentation information, separating the text into articles and adverts). Similarly, the National Library of Finland (http://digi.kansalliskirjasto.fi/), while providing the METS files, explicitly removed scanning information from them until requested otherwise.

These examples highlight how little thought is given to the material dimension of the newspapers in most digital processing pipelines even before the user interface layer. Luckily, the ALTO files of the National Library of Finland had a MeasurementUnit of mm10. Given this, we could easily extract page size, printed area and character and words counts for each page. Besides these, the ALTO file also contains some style information that can be extracted. Currently, we disregard the information on left/center/right alignment, but do extract font information. Directly given are the size, face, style (bold/italic/underline) of each font used, to which we add the calculated number of characters and words written using that font, as well as the overall page area covered.

For each page, we also extract all text box coordinates (visualised in Figure 1). While these are primarily meant to locate text visually on the page in reader interfaces, they can be processed to yield layout information. First, we extract column counts using a lighter-weight process than the computer vision approach used in [6]. We scan the page from top to bottom, for each Y coordinate counting the number of text boxes present there. This yields a distribution associating all column counts with the area they control on the page. Mapping shifts in the amount of columns seems to be one of the clearer indicators of changes in layout. This is useful both for assessing the general development of newspaper layout, but also for identifying particular instances in which editors felt they needed to introduce changes to the layout. Columns obviously roughly correspond to page

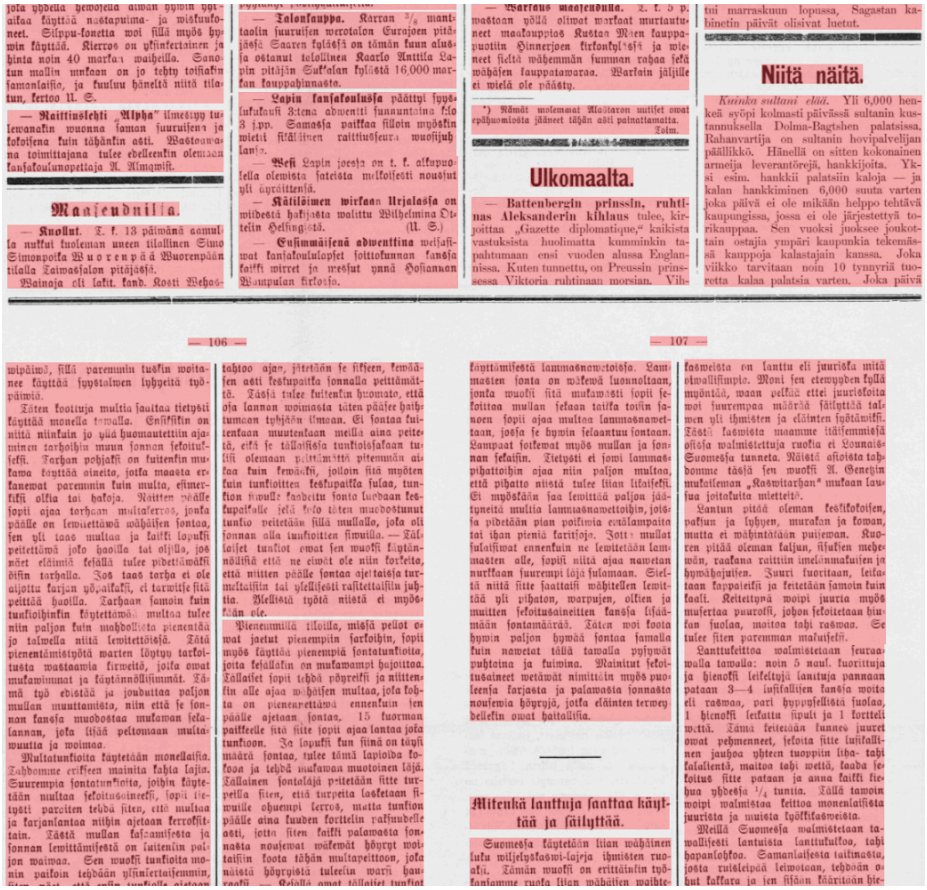

Fig. 1. ALTO text blocks overlaid on a newspaper page.

size, but changes in the width of columns are also indicative of how newspapers explored issues of readability.

2.2 Developing the Materiality Explorer

The Helsinki Computational History Group sits along the same corridor at the University of Helsinki. This physical presence is an important part of the group's work, but so is Slack. As a tool, Slack is an effective way of communicating while sharing research ideas and findings, but it also has the benefit of functioning as a means of documenting much of the group's efforts. To provide an example of this, we will present shortly below an analysis of our Slack communication relating to developing the materiality explorer.

On this particular project, the intensive work started – according to the comments on Slack – on 30 October 2018. It began when Eetu Mäkelä posted first images of a general visualisation unifying multiple aspects of materiality data. From the beginning, it was clear that the point of the materiality explorer was to experiment with different ways to define gross materiality categories in newspapers. It took however few days before the work on the development got going seriously.

Nevertheless, by 12 December 2018, there were altogether 355 different messages (8-9 on average / day) on the group's slack channel dedicated to newspapers about this work. Altogether 9 people participated in this online discussion with different kinds of input. While some people just posted one or a few notes, two group members had more than 100 messages each devoted to this project. There was also, of course, actual human interaction in real life, which is unfortunately not recorded. What drove the work was a looming deadline for the DH2019 conference at the end of November.

Analyses undertaken on development versions of the materiality explorer soon led us to realise that some of our data was problematic. Here, an important point to notice is that computational processing of the data did not start with us, but included also the scanning and OCR of the pages, as well as the metadata work done on the collection at the National Library of Finland. What we found out was that the National Library of Finland had used altogether 22(!) versions of scanning software. A key problem for us was that some of these did not differentiate between Fraktur and Antiqua fonts. By using metadata to analyse which newspapers were scanned with which version, we determined that reliable font identification could only be had up to the year 1910. We also employed some spot checking to compare algorithmic results to the manually keyed metadata, and for example decided to use the raw data directly for page size and date range estimation instead of the same information as keyed.

After a few days of pondering about the effects of these technical problems for analysis, we started focusing more on the question of cramming information on one sheet of newspaper – thinking also about the readability of the text on the page. At the same time, a more extensive reading of relevant secondary sources begun to figure out the technological development (especially with the DH2019 conference submission in mind). The reason for doing this was to find possible

identifiable markers to flag differences and effects caused by changes in printing techniques. For example, the emergence of lithography offset printing was one such technique whose effects we could clearly identify also in the data.

We also soon advanced to thinking about layout and the relevance of the front-page. The idea was to figure out ways of detecting typographic changes on the front page within the context of a single newspaper to understand its development. At this time, it came as an idea to try to identify an instance of a (statistically) typical front page for each decade over time for both Finnish and Swedish language newspapers. Once we knew that this is possible based on the tools at hand, several different kinds of experiments to find "typical" newspaper proportions using the materiality explorer were made. Our deliberations particularly echoed those by Myllyntaus [21], who has done a huge amount of work on these issues without the statistical apparatus that we have on hand today. What was visible in our data was that importing the rotary press and offsetting technology to Finland changed the newspaper layout in the papers that could afford this technology in a very short period of time. We were able also to see that the linguistic and geographic diversity in Finland led to a situation where print runs were smaller and there was more type-setting ongoing than in some larger European countries.

We realised also that we could group different language newspaper published by the same publisher in the same year at the same location together in order to study their layout and content. This would help us to understand how news possibly circulated from one language to another and how different advertisements for example are presented in different languages in Finland. Many previous scholars have been interested about different language profiles of newspapers in different Finnish towns. What these scholars haven't realised is that the question of type, layout etc. can also have intellectual relevance. So, to ask if parallel newspapers are coming from the same publishing house (as they at times do) is a relevant question to ask.

On Sunday 25th of November, Eetu Mäkelä posted an image of the mean front page of Helsingin Sanomat in 1907. This also marked the saturation point of the development phase of this part of the work. There were still new ideas coming in, for example, about terseness of language in newspapers in order to allow cramming, but the main thing for us at this point was to prepare for the DH2019 deadline that was on 27th of November. Perhaps we need to wait for the next deadline to get back seriously to this project.

2.3 The Materiality Explorer Interface

As it currently stands, the materiality explorer has three main functionalities, each aimed at a different use cases. Common to all views are a set of selectors, allowing to limit the set of newspapers under study. Currently, these hold facilities for limiting study by 1) time, 2) newspaper language, 3) newspaper lifetime, 4) printing location and 5) individually by title.

In the overview view shown in Figure 2, first presented is the absolute amount of data. This is important, as all the other graphs display their information as

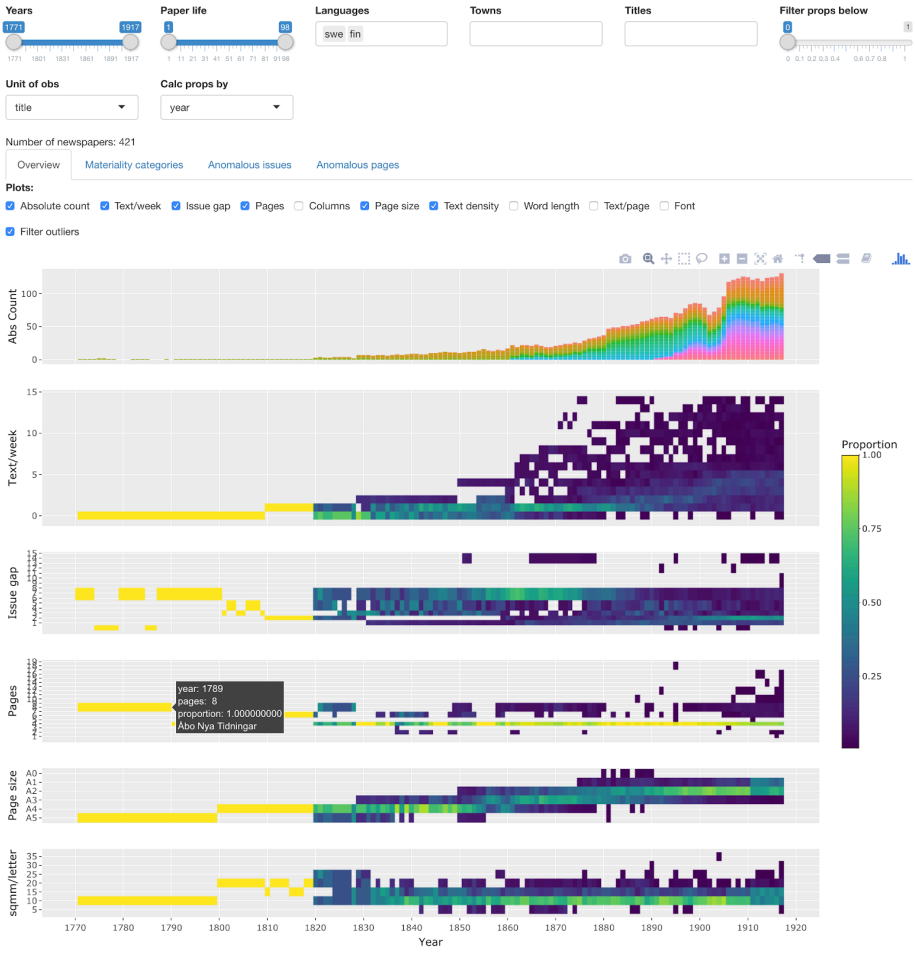

Fig. 2. The materiality explorer.

proportions of the whole. Depending on a user selected option, this proportion may be calculated by year, by month or by week. In addition, the user can select whether they want an observation to be titles, issues or pages. Here, the choice depends on what one is interested in. Counting by titles treats each newspaper as a single unit, allowing exploration of the breadth of newspapers without regard to how often they appeared or how large they were. On the other hand, if one is more interested in the amount of information consumed by an end reader, then possibly counting by issue or even by page is appropriate. Another use case where observing by page or issue may be more interesting is when studying the development of a single newspaper, where the differing publication frequencies and page sizes no longer matter, but instead even singular aberrant pages are interesting.

After this absolute count, a baseline measure of text per month is given, against which all the materiality information can be contrasted. This baseline was developed in consultation between the computer scientists, historians and linguists to provide a language-neutral measure for throughput. By counting the number of characters each newspaper produces in a month without regard to how they are divided between issues or pages, this measure shows how much content needs to be transmitted. As this quantity rises, newspapers must respond with material innovations, whether by increasing page count, page size or publication frequency, or by cramming more material into available space by decreasing font size, line breaks or margins.

A second view allows grouping the data by a combination of material dimensions, thereby allowing exploration of archetypal materiality categories. Finally, two distinct views allow the user to explore respectively page and issue-level material anomalies in the data: for example pages which have much more text than others or pages with abnormal layout, or issues with appendixes or which appear on the same day as another issue. These both lead the way for interesting qualitative analyses, but can also be used to remove abnormal data from further quantitative computational analyses of either form or content. In our project, the anomaly detection served as a central method for exploring the data as well as identifying errors in the code or metadata. Here historians had a rich source for detecting counter-intuitive findings, and often those findings led to feedback that could further improve coding efforts.

At present, we are using the interface to exploratively develop hypotheses on common development patterns as well as archetypal materiality categories. Both of these are interesting in themselves as objects of study, but can also be used later to partition datasets for other computational processing such as OCR retraining or content analysis. While this current stage of explorative hypothesis development is interactive, visual and qualitative, our plan for the next stage is to explore statistical validation of such hypotheses using for example Granger causality and archetypal coverage measures. Once developed and tested, these again will be added to the interface to enable further self-sufficient analysis in a more trustworthy manner.

3 Discussion

At the outset of this project, we asked in particular how the modernisation of newspapers published in Finland could be better understood by looking at the form, shape, location and publication frequency in newspapers published in different languages (Finnish and Swedish being the main publication languages). The project produced one article that pays particular attention to the different speed in development with regard to Swedish-language and Finnish-language newspapers in Finland. Further, we produced an interactive materiality explorer that helps researchers understand the development of material aspects of newspapers. We also developed preliminary hypotheses that will be shortly discussed below with regard to different categories of materiality.

For the Finnish newspapers, the data shows a general order in how they expanded: first, layout was changed to include more words per page; second, page size was increased; third, publication frequency was increased and only after that was the amount of pages increased. This last step often coincides with the introduction of rotary presses, which allowed newspapers to more easily be composed of more than four pages, and also allowed them to move back from large page sizes to more easily handled formats. Simultaneously, the data shows also high variability, where papers not only frequently printed supplements, but could switch back and forth between formats inside a single week, or cram text into a special issue through diminished line breaks. Similar shifts took place also with regard to fonts. Newspapers explored different Fraktur and Antiqua fonts to try out readability, but also because fonts were oftentimes used to signal that the contents was aimed for a particular audience. While there are plenty of exceptions to this, it seems that Fraktur was more often used when dealing with economy and religion, whereas Antiqua was reserved to politics, philosophy and the high arts. To test such hypotheses about different uses for fonts and relating that to the overall development of newspapers, we still need more robust statistical information. We also aim to compare used fonts and with other factors, such as language frequency and size of newspapers. (For the history of newspaper layout and design, see [4,17,22,24,26,27,12].) Compared to earlier studies, our data driven approach gives us a great opportunity to evaluate the main findings of earlier historical studies of newspaper materiality [18,30,21,9,16,32].

What we also aim to do with these patterns is to develop evidence-based archetypal categories of newspapers across history. We are then able to trace and compare these through time and place, but also use them to study the evolution of individual newspapers. These categories will also help us understand the newspapers as objects of intellectual activity, creating a theory of different historical maturity levels of newspapers. This in turn will help us chart the development of public discourse over time.

Besides presenting the research process regarding the material development of newspapers as a genre in itself, we argue that content and form interact, and thus big data approaches to newspaper analyses also need to pay attention to material differences in order to accurately understand the subdivisions in large corpora. Here, this paper continues on a path previously charted by for example

[19,29,14], while providing an orthogonal axis to those expanding study from text to visual elements [25,31]. For example, using the metadata we can create meaningful subsets of the data that are balanced by paper type for for example topic modelling or teaching automated transcription algorithms.

Here, Finland makes an intriguing case for digital history because its public sphere is bilingual, with newspapers in both Swedish and Finnish. One interesting phenomena that arises from this are publishers publishing newspapers in both languages. For example, in Kotka there are both Finnish and Swedish newspapers by the same publisher with identical layouts and advertisements. Such could be used to create parallel corpora, interesting for the study of commonalities and differences between the different language public spheres, but also perhaps as material for machine translation.

References

1. Allen, J.E.: The Modern Newspaper: its typography and methods of news presentation. [With illustrations. New York & London, Pp. ix. 234. Harper & Bros. (1940)
2. Baldasty, G.J.: Commercialization of News in the Nineteenth Century. University of Wisconsin Press, Madison (2014)
3. Bode, K.: The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History. Modern Language Quarterly **78**(1), 77–106 (Mar 2017). https://doi.org/10.1215/00267929-3699787, https://read.dukeupress.edu/modern-language-quarterly/article/78/1/77-106/19924
4. Broersma, M. (ed.): Form and style in journalism: European newspapers and the representation of news 1880-2005. Peeters, Leuven, Dudley, MA (2007)
5. Buntinx, V., Bornet, C., Kaplan, F.: Studying Linguistic Changes over 200 Years of Newspapers through Resilient Words Analysis. Frontiers in Digital Humanities **4** (2017). https://doi.org/10.3389/fdigh.2017.00002
6. Buntinx, V., Kaplan, F., Xanthos, A.: Layout analysis on newspaper archives. In: DH2017 abstracts (2017), https://dh2017.adho.org/abstracts/193/193.pdf
7. Cordell, R., Smith, D.: What News is New?: Ads, Extras, and Viral Texts on the Nineteenth-Century Newspaper Page. In: DH2017 abstracts (2017)
8. Cristianini, N., Lansdall-Welfare, T., Dato, G.: Large-scale content analysis of historical newspapers in the town of Gorizia 1873–1914. Historical Methods: A Journal of Quantitative and Interdisciplinary History **51**(3), 139–164 (Jul 2018). https://doi.org/10.1080/01615440.2018.1443862
9. Gustafsson, K.E., Rydén, P.: A History of the Press in Sweden. NORDICOM, Göteborg (2010)
10. Hutt, A.: The changing newspaper; typographic trends in Britain and America 1622-1972. Gordon Fraser, London (1973)
11. Høyer, S., Pöttker, H.: Diffusion of the news paradigm 1850-2000 (2014)
12. Kapr, A., Forssman, F., Willberg, H.P.: Fraktur: Form und Geschichte der gebrochenen Schriften. Hermann Schmidt, Mainz (1993)
13. Kutsch, A.: Journalismus als Profession: Überlegungen zum Beginn des journalistischen Professionalisierungsprozesses in Deutschland am Anfang des 20. Jahrhunderts. In: Blume, A., Böning, H. (eds.) Presse und Geschichte: Leistungen und Perspektiven der historischen Presseforschung, pp. 289–325 (2008)

14. Lahti, L., Marjanen, J., Roivainen, H., Tolonen, M.: Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly **0**(0), 1–19 (Jan 2019). https://doi.org/10.1080/01639374.2018.1543747
15. Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., FindMyPast Newspaper Team, Cristianini, N.: Content analysis of 150 years of British periodicals. Proceedings of the National Academy of Sciences **114**(4), E457–E465 (Jan 2017). https://doi.org/10.1073/pnas.1606380114
16. McReynolds, L.: The News under Russia's Old Regime: The Development of a Mass-Circulation Press. Princeton University Press (1991). https://doi.org/10.2307/j.ctt7zth51
17. Moen, D.R.: Newspaper layout and design. Iowa State University Press, Ames (1989)
18. Moran, J.: Printing Presses: History and development from the Fifteenth Century to Modern Times. University of California Press (1973)
19. Moreux, J.P.: Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment (Aug 2016), https://hal-bnf.archives-ouvertes.fr/hal-01389455/document
20. Morison, S.: The English Newspaper, 1622-1932: An Account of the Physical Development of Journals Printed in London. Cambridge University Press, 1 edition edn. (Oct 2009)
21. Myllyntaus, T.: Suomen graafisen teollisuuden kasvu 1860-1905. University of Helsinki, Helsinki (1981)
22. Olson, K.E.: Typography and Mechanics of the Newspaper. D. Appleton and Company (1940)
23. Pettegree, A.: The Invention of News: How the World Came to Know About Itself. Yale University Press (2014)
24. Presbrey, F.: The history and development of advertising. Doubleday, Garden City, N.Y. (1929)
25. Smits, T.: Illustrations to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860-1940. In: DH2017 abstracts (2017)
26. Sutton, A.A.: Design and makeup of the newspaper. Prentice-Hall (1948)
27. Swanson, G.: Graphic Design & Reading: Explorations of an Uneasy Relationship. Allworth Press (2000)
28. Tilles, D.: The Use of Quantitative Analysis of Digitised Newspapers to Challenge Established Historical Narratives. Roczniki Kulturoznawcze **7**(1), 83–97 (2016). https://doi.org/10.18290/rkult.2016.7.1-4
29. Tolonen, M., Lahti, L., Roivainen, H., Marjanen, J.: A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History **0**(0), 1–22 (Dec 2018). https://doi.org/10.1080/01615440.2018.1526657
30. Tommila, P., Salokangas, R.: Sanomia kaikille: Suomen lehdistön historia. Kleio ja nykypäivä, Edita, Helsinki (1998)
31. Wevers, M., Smits, T., Impett, L.: Modeling the Genealogy of Imagetexts: Studying Images and Texts in Conjunction using Computational Methods – DH2018. In: DH2018 abstracts (2018), https://dh2018.adho.org/en/modeling-the-genealogy-of-imagetexts-studying-images-and-texts-in-conjunction-using-computational-methods/
32. Wilke, J.: Belated modernization: form and style in German journalism 1880-1980. In: Broersma, M. (ed.) Form and style in journalism : European newspapers and the presentation of news, 1880-2005, Groningen studies in cultural change, vol. 26. Peeters, Leuven, Dudley, MA (2007)

Twinning Classics and A.I.: Building the new generation of ontology-based lexicographical tools and resources for Humanists on the Semantic Web

Maria Papadopoulou[1,2] and Christophe Roche[1,2]

[1] University Savoie Mont-Blanc, France
[2] Liaocheng University, China
firstname.lastname@univ-smb.fr

Abstract. This Twin Talk is about the ongoing collaboration between an expert in Classics and an expert in Artificial Intelligence (A.I.). Our approach set out to answer two interlinked issues, ubiquitous in the study of material culture: first, pairing things to their names (designations) and, second, having access to multilingual digital resources that provide information on things and their designations. Our chosen domain of application was ancient Greek dress, an iconic feature of ancient Greek culture offering a privileged window into the Greek belief systems and societal values. Our goal was to place the Humanist/domain expert at the centre of the endeavour enabling her to build the formal domain ontology, without requiring the assistance of an ontology engineer. The role of A.I. was to provide automations that lower the cognitive load for users unfamiliar with knowledge modelling. Building the model consisted in distinguishing between concept level (i.e. the stable domain knowledge) and term level (i.e. the terms that name the concepts in different natural languages), putting these into relation (i.e. linking the terms in different languages to their denoted concepts), and providing complete and consistent definitions for concepts (in formal language) and terms (in natural language).

Keywords: ancient Greek dress, ontology, terminology.

1 Introduction

The proposed Twin Talk is about the story of an interdisciplinary collaboration between an expert in the Humanities (Classics) and an expert in digital technology (Artificial Intelligence) working together to answer two interlinked issues, ubiquitous in the study of material culture, broadly defined as "the investigation of the relationship between people and things irrespective of time and space" [1]: first, pairing things to their names (designations) and, second, having access to multilingual digital resources that provide information on things and their designations. Our chosen domain was ancient Greek dress, an iconic feature of ancient Greek culture which offers a privileged window into the Greek belief systems and societal values. The challenge was triple:

a/ deal with the complex history of terms designating ancient Greek dress, some inherited from ancient times, others coined by scholarship dating since the Renaissance [2-3].
b/ define concepts formally, yet in a way that would be intuitive to the Humanist-Classical scholar, enabling her to do the ontological modelling on her own.
c/ model domain concepts *and* terms providing definitions for both (i.e., formal definitions for concepts; natural language definitions for terms, based on the concept designated by each term).

The paper is organized in five sections: Section 1 is the Introduction. Section 2 presents the chronicle of this ongoing collaboration. Section 3 introduces the problem addressed and explicates the solution given and the reasons for choosing it. Section 4 provides record of the fruit of the collaboration. Section 5 relates the particulars of the collaboration experience. Finally, section 6 reports on the lessons learnt and suggests a number of good practices.

2 Teaming up: the Classical scholar and the Artificial Intelligence (A.I.) expert

Our team of two is made up of researchers at different career stages and with different academic backgrounds, coming from disciplines as disparate as Classics and A.I. It was initially formed with the aim to combine our diverse expertise in lexicography, classics and dress studies (Classicist), terminology, ontology and A.I. (digital expert) to model knowledge and terminology used to express this knowledge in the domain of ancient Greek dress. Our working hypotheses are as follows: a/ while ancient Greek garments and their names are culture-specific, ancient Greek dress concepts can be described in a context-free, formal manner that enables sharing them across different natural languages; b/ concepts are defined by a set of *essential characteristic* known to domain experts, traceable in texts, and visible through the representations of dress in sculptures, painted vases, coins, etc. (a characteristic is *essential* for an object iff, when removed from the object, the object is no more what it *is*. For example, 'without sleeves' is an essential characteristic for an exomis)

We first met at the 2013 Terminology and Ontology: Theories and Application (TOTh) workshop organized at the University of Copenhagen, where we both contributed papers [4]. This was a happy coincidence that kick-started a series of academic exchanges: the Humanist expressed the grave problem she encountered when dealing with the terminology of dress in her domain. The Digital expert promised that this was feasible and started to explain why. Soon they realized that this problem was part of a wider problem facing the whole community that worked with textile and dress terminology, which seemed unsolvable for decades: textiles and dress scholars need to standardize the language used in order to communicate knowledge about the objects of the domain, so that everybody understands the same thing.

After the launching of the Humanist's Marie Curie Fellowship at the University of Copenhagen (2015-2017), the team started working on modelling the domain of ancient Greek dress. This work was a deliverable of the Humanist's two year project, as shown in the final project report [5]. In January 2017 the Humanist joined team Condillac-LISTIC lab, at the University of Savoie, France. Condillac was founded by the Digital expert several years back [6]. It is an international and interdisciplinary team primarily of computer scientists and linguists working on Knowledge Representation. In November 2017 a new research center was opened at the Computer Department of Liaocheng University, China [7] and the Digital expert asked the Humanist to present their work and give lectures on Digital Humanities in China, so that more students and researchers in Computer Science as well as in the Arts and Humanities would become familiar with the idea of embarking on digital humanities projects individually or in teams made up of a humanist and a computer scientist. The two researchers' collaboration is ongoing in the context of both Condillac and KETRC.

3 Problem and solution

3.1 The problem of 'naming things' in the experts' own words: a terminology and knowledge modeling issue

Scholars vividly express a need, omnipresent in the study of material culture terminology, whether the research area is ancient Greek dress (cf. d, e, h, j), dress of medieval Scandinavia (cf. c), Greek material culture (cf. a), ancient Egyptian art (cf. b, k), clay pottery from different cultures (cf. f, g, i) to:

i) Determine what term goes with what object combining textual, iconographic, material sources:

a. "Only studies that combine archaeological and iconographic data with knowledge derived from texts give the opportunity to correlate a word with an object." [8]

b. "In our case, the content of the terminology pertains to two fields: objects and pictures. When saying "objects", I mean the entire material culture ... the terms we are looking for pertain, obviously not to the specimens existing in reality, but to every occurring type of objects and buildings. We have to do here with a list of designations of things." [9]

c. "Research into dress history, whether the approach is founded in history, art or archaeology, incorporates terminology, one way or another." [10]

ii) Adopt a standard common vocabulary of terms and definitions to promote research in their field:

d. "Although the standard Greek and Latin terminology employed by scholars to describe ancient clothing may not be that which was used in antiquity … it is a useful vocabulary of dress and will be used here." [13]

e. "Studies of garment-terms in historical societies tend to be hampered by a lack of understanding of the specific vocabulary of dress." [11]

f. "…it would help if we could work out a list of standard vessel shapes, clearly defined and illustrated, and a set of terms for them." [12]

g. "An intelligent discussion of pottery shapes is rendered more difficult by lack of definitive nomenclature." [13]

h. "… Arabic terms for specific veil types (words like shaal, maghmuq, and lithma) … will be used to identify certain ancient Greek veil-styles. This might not be the most satisfactory answer, but at least it is expedient: we need to adopt a common workable veil-vocabulary so that our investigation of the Greek veil can proceed without further complication or impediment." [11]

iii) Have access to diachronic multilingual resources providing information on things and their names:

i. "…that we seek standardized terms for ceramic vessels expresses what I feel to be a real need…develop multilingual vocabularies of technical terms" [14]

j. "Creating a diachronic and global costume term base…is of considerable value for textile terminology." [15]

k. The chief aim of a terminology is efficient communication among specialists when discussing matters orally or in written form, efficient organisation of data banks, and - a point of particular importance - successful communication among electronic data banks. [9]

Dress scholars have attempted to unravel the complexity of dress terminology [17-18] and produce a classification of clothing in order to meet the need for a transcultural denomination system for clothing parts, but the domain of ancient Greek dress has never been described using a stable vocabulary [19-22]. Contemporary needs for machine tractable data on the Web impel the use of software artefacts, i.e., controlled vocabularies, thesauri, ontologies, to structure domain knowledge. Yet, existing thesauri, i.e., the Getty Art and Architecture Thesaurus (AAT) [23], ontologies of the domain of dress [24], the CIDOC CRM, which provides definitions and a formal structure for describing concepts and relationships in cultural heritage documentation [25], do not cover the needs of scholars interested in ancient Greek dress, as they do not include any Greek terms, apart from *chlamys(es)*, *chiton(s)*, *peplos(es)* and *himation(s)* in AAT.

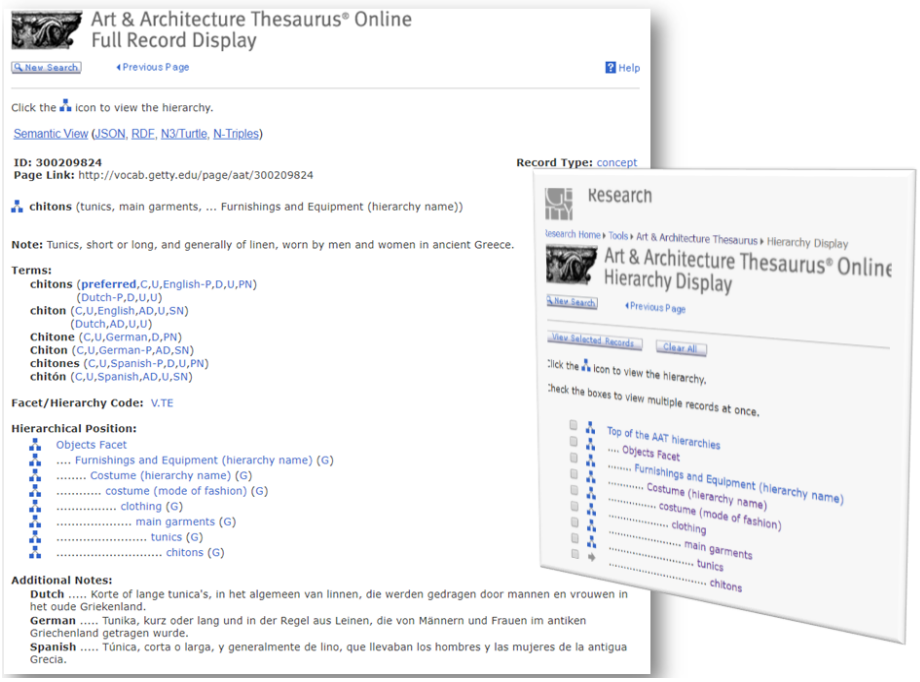

Fig. 1: Art & Architecture Thesaurus *s.v. chitons*

3.2 The solution: empowering the Classicist

Our aim was first, to match the names (terms) to the objects (concepts) of the domain by defining them with consistency, then, structure and publish these terminological data as shareable and reusable Linked Open Data with the help of a software platform that would build the concept system based on defining concepts as sets of essential characteristics, in compliance to ISO 108) [16]. We wished to achieve the above tasks using workflows and tools that empower classicists and humanists by matching their way of thinking and working, not the way of thinking and working of Semantic Web experts and developers. The domain of application was ancient Greek dress and its culture-specific terminology. Its importance as a social marker or as representative of the materials, techniques and technological know-how of a given era is unquestionable. Unraveling the intricacies of Greek dress terms, building the concept system of this domain, and publishing both terms and concepts as Linked Open Data, is to be the first step towards making this knowledge easily discoverable and reusable.

Our ontological modelling is informed by a theory of concept inspired by the international standards on terminology [16, 26]. According to ISO 1087 [16] concept is a "unit of knowledge created by a unique combination of characteristics"; characteristic is an "abstraction of a property of an object or of a set of objects; essential characteristic is a characteristic "indispensable to understanding a concept". Identify and define what

the objects and what the terms that designate them are, are constant ontological concerns in both textual and object-based research of Greek antiquity. The inclination and capability to classify, sub-classify and use the right terms is the dividing line between an expert and a lay person. As we have already shown [27-28], experts in ancient Greek dress and other domains of cultural heritage regret the "terminological vagueness" that prevails in their respective fields, blame it for hindering their work, and emphasize the need for consistent and consensual use of terms. As archaeological finds come down to us like a picture book without names, and as texts furnish names for objects, without providing illustrations, deciding "what was what' is not easy. Names, what the discipline of Terminology calls "terms", provide shortcuts to communicating what things *are*. Terms are words that belong in a domain-specific language, not general language [29]. Terms mediate between concepts and language. A term cannot exist without a concept, while a concept may have a verbal expression in Language A, but not in Language B. Concepts are abstracted from individual objects. They are the layer that mediates between 'reality', where objects live, and language, where terms communicate meaning about concepts. Concepts are bits of knowledge about the world, while terms are bits of language to verbally express this knowledge.

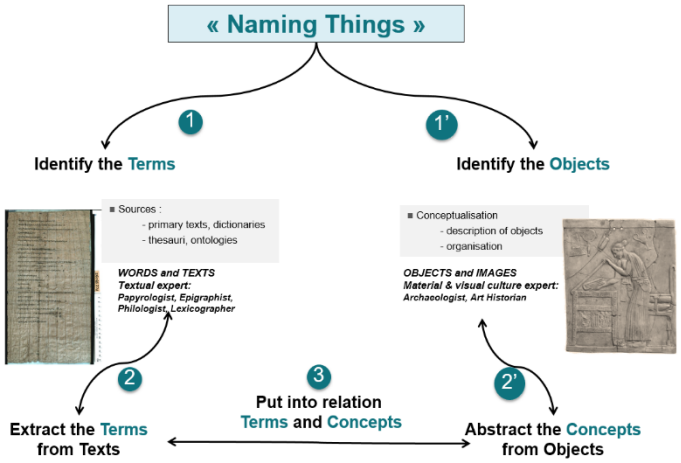

Fig. 2: Naming Things in the domain of ancient Greek dress

The move from linguistic-textual discursive to extra-linguistic meaning is important, when defining extra-linguistic entities, i.e., concepts and objects [30-32]. Our proposed solution for the domain of Greek dress, but also for experts seeking to "name things" in their respective domain, is to define the terms of the domain in relation to the concepts designated by these terms and build their term list in connection to the ontology made up by a concept *isa* hierarchy, whereby each concept (and term labelling this concept in a natural language) is made up of a list of characteristics, following the ISO 1087 standard for terminology work.

The role of Artificial Intelligence (A.I.) in this endeavour is to reduce the complexity of representing concept characteristics, concepts and terms and provide helpful automations to bridge the gap between classicists/humanists and the Semantic Web, while promoting two good practices for developing methodologies and tools in Digital Humanities: *intuitiveness* (of the theoretical and methodological framework to be devel-

oped), *ease-of-use* (of the software platform to be developed). To illustrate the complexity of modelling: i.e. 10 differences leading to forming disjoint categories (e.g. with or without sleeves), suffice to end up with a Porphyry tree of 1024 (2^{10}) terminal concepts. Selecting a terminal concept out of this complexity requires the aid of the machine.

Ontologies (in Computer Science) have been around for the last forty years or so [33-34] and OWL ontologies have been around since 2004. They are the best tool to describe domain knowledge, publish metadata compliant with Semantic Web and Linked Data standards, annotate resources, and query knowledge bases; they are the backbone of the Semantic Web [35]. But the standard language for ontologies in the Semantic Web is OWL (Web Ontology Language) [36]. Modelling in OWL using Protégé [37-38] (or another platform for editing [39]) requires reasoning in Description Logics.

The reasons we opted out of building an OWL ontology in Protégé are both epistemological & practical: first, reasoning in OWL using Protégé means reasoning in role restrictions, classes and individuals, data properties, object properties, A-box and T-box, which is hardly intuitive to those with no background in Description Logics and does not match the way classicists/humanists work. To do the modelling in this way, domain experts either need an ontology engineer, or have to think like one. Second, research has shown that human users do not fare well with highly formal systems, unless they have background in Computer science [40-42]. The Semantic Web is based on logical reasoning (first order logic), which requires a highly degree of formalization. The use of a formal language with clearly specified syntax and semantics, such as Description Logics at the heart of Semantic Web ontologies guarantees the consistency of definitions and the possibility to reason on these models, but is not consensus-oriented. It is much more intuitive to humanists to define the objects of the domain in terms of knowledge primitives that can be traced

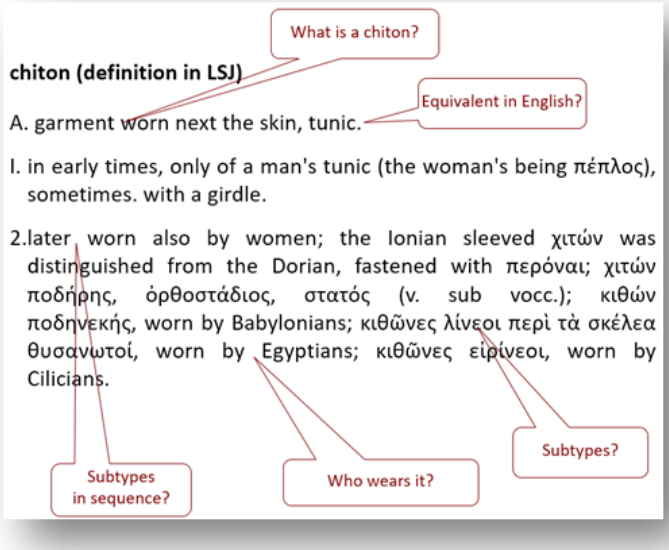

Fig. 3: Traceable knowledge primitives in the LSJ definition of *chiton*

in dictionary definitions, primary texts & archaeological objects. Fig. 3-4 illustrate the

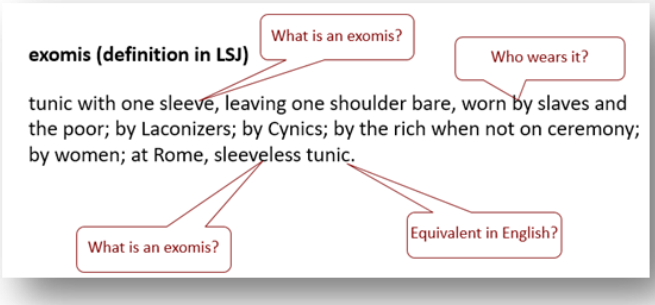

Fig. 4: Traceable knowledge primitives in the LSJ definition of *exomis*

knowledge primitives of *chiton* and *exomis* respectively traceable in its definition in LSJ [43-44, the standard dictionary used by classical scholars. These are: worn next to the skin, (initially) by men, (later also) by women. Additional characteristics subdividing this type of Greek garment into subtypes are included in the LSJ definition. Knowledge primitives can be a firm basis for consensus-reaching discussions among domain experts. Domain experts should be given common ground for agreeing (or disagreeing) on the definitions of terms and concepts. Description Logics does not guarantee a common basis upon which a dialogue among experts can exist. In contrast, semantic primitives of concepts can form a stable basis for scholarly discussions on the meaning of concepts and terms.

In order to build our domain ontology, we used Tedi [45], a software developed by the digital expert, which empowers domain experts. Tedi software supports both term standardization and customization. Standardization of terminologies relies upon expert agreement on domain knowledge, which is necessary for collaboration and rapid sharing of information. Customization accommodates and preserves the diversity of terms across languages. Tedi's complex architecture deploys two interconnected levels:

- the formal domain ontology level, which consists of an editor for concepts and an editor for objects. The editors of attributes, relations, and axes of analysis are accessible by means of the concept editor.
- the terminology level, which consists of an editor for terms and an editor of proper names.

For the user's convenience the interfaces are color coded: green for the conceptual dimension, blue for the linguistic dimension. Tedi allows ontoterminologies to be exported in different formats human readable, as well as machine tractable and Semantic Web compliant: HTML (static and dynamic), RDF/OWL, SKOS, JSON, and CSV.

3.3 A new scholarly workflow for building definitions for things

The Tedi tool-based method for building multilingual ontoterminologies is composed of 5 interrelated tasks, which do not necessarily have to be performed in a linear

fashion. The first step is to define the concepts of this complex domain in a formal language by means of specific axes of analysis. The next step is to associate each term with the concept made of the chunks of knowledge essential to defining it. Such modelling can structure knowledge so as to eventually support two types of queries: by means of keywords, and by means of concepts. Fig. 5 illustrates the linking of concepts to terms by means of selecting essential characteristics.

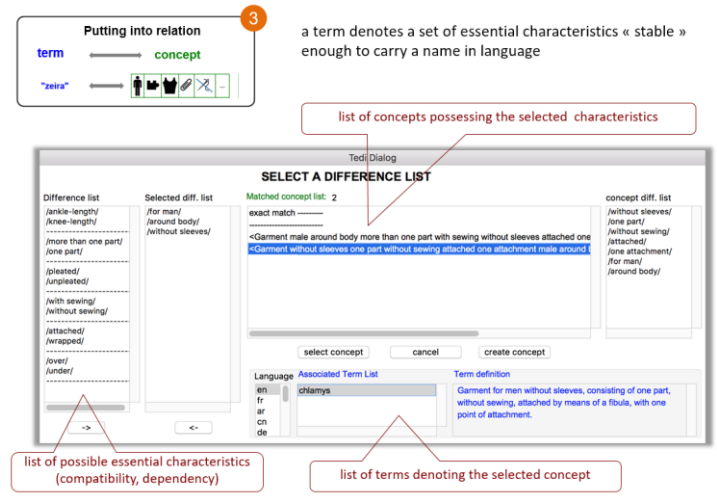

Fig. 5: Selecting the essential characteristics for *exomis*

The ontology has led to the building of an ontology-based online dictionary, whose definitions of term were definitions of thing. Using the example of the *exomis*, we have arrived at the following definitions, in English, French and Modern Greek:
“exomis” : Short and non-pleated garment for man, usually worn around the body directly on the skin, this sleeveless garment consists of two pieces of cloth sewn together along the sides, attached on the left shoulder leaving the right shoulder and part of the chest naked.

“Exomide” : Vêtement de corps pour homme, court, non-plissé et sans manches. Composé de deux pièces cousues le long des côtés, attaché sur l’épaule gauche laissant l’épaule droite et une partie de la poitrine nues, il est généralement porté directement sur la peau.

“Εξωμίδα” : Κοντό, χωρίς πτυχώσεις και χωρίς μανίκια ανδρικό ένδυμα, το οποίο συνήθως φοριόταν ως κυρίως ένδυμα. Αποτελούνταν από δύο κομμάτια υφάσματος ραμμένα στα πλάγια και στερεωμένα στον αριστερό ώμο που άφηναν τον δεξί ώμο καθώς και μέρος του στήθους ακάλυπτα.

The LSJ, the bilingual Greek-English dictionary commonly used in the field of classical studies, defines exomis as a “tunic with one sleeve”. This definition is not only incomplete, but also problematic with regard to the notion of “sleeve”: “The adjective amphimaschalos attributed to the Greek chiton in no way implies the idea of sleeves, but only, by its very etymology, that of the two armpits ... it is abusively, in my opinion,

that our translators or lexicographers speak of ‘sleeve’ tunics when it is a tunic with two armholes.” [46, our translation].

This approach led to building thing definitions, i.e. definitions of the concept denoted by the term, and was not aimed at representing term meanings in discourse. The result is precise and complete formalized knowledge allowing to verify logical properties for multilingual semantic searches and semantic annotations. The objective of our approach is not to impose definitions, but to propose definitions (in natural language) that are based on domain knowledge. This approach allows experts to discuss objectively on the basis of the essential characteristics on which they generally agree. Fig. 6-7 show the definition for *exomis* as exported in fully human readable and machine processible exports (dynamic HTML and OWL respectively). The Tedi Onto-Dictionary of terms and concepts will be deposited in Clarin.

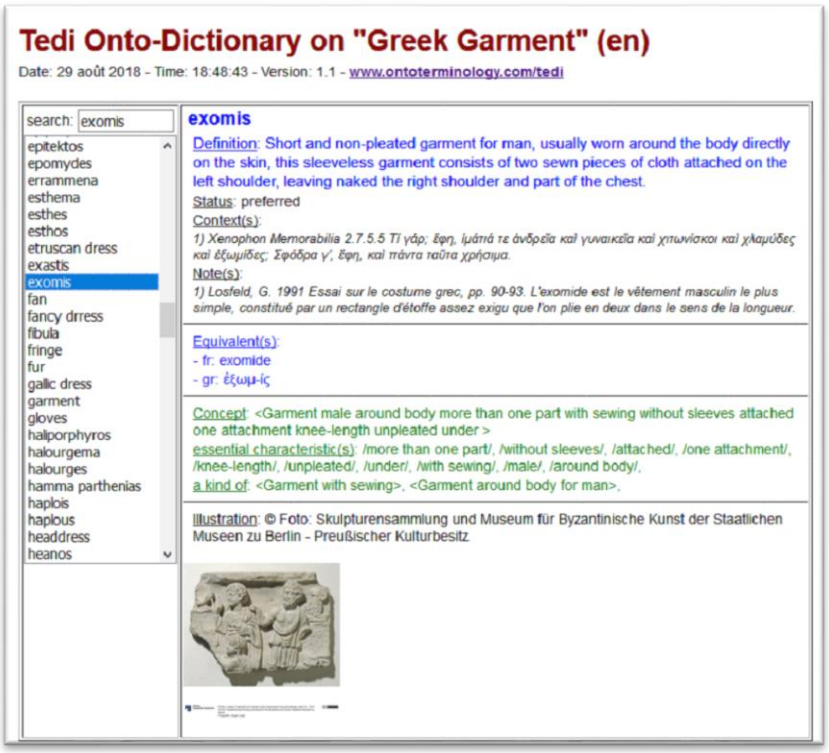

Fig. 6: Tedi export of the entry for term “exomis” in dynamic HTML

Fig. 7: A fragment of the ontology in OWL

4 Academic output & other achievements

The model for collaboration between a classical scholar and an Artificial Intelligence expert has numerous achievements to show for:

- The Onto-dictionary of ancient Greek dress;
- A new software for onto-terminologies standalone and in a proprietary language;
- A more recent attempt to model the domain of ancient Greek vases [47];
- Disseminating the idea of interdisciplinary collaboration between Humanists and Digital experts by means of Condillac-LISTIC (France) and KETRC (China).
- Testing their idea for ontological modelling of terminologies from Humanities' disciplines with colleagues from different communities (in international conferences and peer-reviewed journals) including a best paper award [51]: archaeologists (Institut français d'archéologie orientale-Cairo 30-31 October 2016), digital classicists and digital humanists (European Association for Digital Humanities 2018) [48], terminologists (TOTh 2016) [27], information scientists-librarians-archivists (AIDAinformazioni Journal) [28], translators-lexicographers-linguists (Lexicologie Terminologie Traduction-LTT) [49], Artificial Intelligence experts (Revue Intelligence Artificielle special issue on DH and AI) [50], computer scientists (several papers given in China [7], Semapro 2018 [51]).

5 The collaboration experience

5.1 The good stuff: a mutually empowering experience

Cambridge English Dictionary defines collaboration as "the situation of two or more people working together to create or achieve the same thing" [52]. Our collaboration flourished thanks to our positive attitude and openness. We agreed on the research goal, specific objectives, approaches, and methodology. Especially because this was a cross-border collaboration, online meetings were scheduled at regular intervals, having specified the details of the agenda beforehand. In terms of accountability, both researchers accepted full responsibility for the actions, as well as to disclose the results in a transparent manner. Our collaboration is informed by the principles laid out in the European Code of Conduct for Research Integrity [53].

5.2 The challenging stuff

In 1993 Turner and Cochrane [54] suggested that there are four types of projects according to two parameters: how well defined their goals are, and how well defined the methods of achieving them are. In our collaborative project the goal was clearer to the classical scholar and the method to the digital expert to start with. Each one had to familiarize oneself with the part which was less clear: the humanist had to cultivate the capacity to operate at a representational level involving types and instances. The digital expert had to adjust to the particularities, uncertainties and gaps in knowledge and information that are common when dealing with past cultures.

6 Suggestions for good practice

The first lesson learnt was that team work is mutually enriching and empowering. The second lesson was that the more one practices interdisciplinary collaborative research, the better one becomes at it. The third lesson is that if a digital solution is offered to Humanists, it should cater to the specific needs of the target community.

Collaboration is common practice among digital humanists. According to a recent study "digital humanities researchers engage regularly in collaborative research. One out of three respondents indicate that they collaborate very often with others on a research project. Altogether, seven out of ten say that they engage often or very often in research collaboration" [55]. If indeed practice makes perfect, digital humanists are well equipped towards setting up collaborations.

Knowledge modelling is interdisciplinary by definition: "In recent years the development of ontologies has been moving from the realm of Artificial-Intelligence laboratories to the desktops of domain experts" [56]. Making cultural heritage term-lists computable in order to link them to other types of resources (e.g., museum objects) is a problem-driven question (as is Ontology Engineering par excellence), not a curiosity-driven one, as in much of the research done in Classics and the Humanities. Our approach aims to show that in order to build workflows and tools that are better suited to the needs of the targeted community, similar interdisciplinary teams are a necessity. We advocate capturing domain knowledge with the help of domain experts, when building ontologies or terminologies whose conceptual system is a formal domain ontology.

How can scholars and digital experts maximize benefits from such collaborations? The answer is to provide training on how to change the way of thinking, i.e. training Computer Scientists on how to think like a Humanist (i.e., a researcher who seeks to understand and analyze how humanity manifests itself in different periods, cultures, media etc.) and train Humanities' scholars on how to think like a Computer scientist (i.e., someone who develops digital tools and media for real-life problem solving). Getting to understand each other's way of thinking raises awareness and improves not only the product, but also the process of the collaboration.

References

1. Editorial: Journal of Material Culture, 1(1), (1996). https://journals.sagepub.com/doi/pdf/10.1177/135918359600100101, last accessed 2019/2/13
2. Lee, M.: The Peplos and the 'Dorian Question'. In: Donohue A. A. and. Fullerton M. D. (eds.), Ancient Art and Its Historiography, pp. 118-147. Cambridge University Press, Cambridge (2003).
3. Stears, K. E. Dress and Textiles. In: E. Bispham, T. Harrison, B. A. Sparkes (Eds.) The Edinburgh Companion to Ancient Greece and Rome, pp. 226-230. Edinburgh: Edinburgh University Press. (2006).
4. TOTh Workshop 2013, http://toth.condillac.org/workshop-2013, last accessed 2019/01/08.
5. Chlamys: The cultural biography of a garment in Hellenistic Egypt Periodic Reporting 1, https://cordis.europa.eu/project/rcn/195523/reporting/en, last accessed 2019/01/08.
6. Condillac Research Group in Knowledge Engineering at the University of Savoie, http://new.condillac.org/, last accessed 2019/01/08.
7. Knowledge Engineering and Terminology Research Centre at the University of Liaocheng, http://ketrc.com/, last accessed 2019/01/08.
8. Ballet, P., J. L. Fournet & M. Mossakowska-Gaubert: Artefacts from Egypt. Archaeology and Texts. Call for Papers. Workshop. IFAO-Cairo (2016).
9. Müller, M.: What was what in ancient Egypt? Terminology development project: art history and iconography. Paper read at the "Egyptologie et informatique" meeting, Würzburg, July 14th, 2000. (2000).
10. Dahl, C.-L.: The Use of Terminology in Medieval Scandinavian Costume History: An Approach to Source-Based Terminology. In: Andersson-Strand E. et al. (eds.) North European Symposium for Archaeological Textiles NESAT X. pp. 41-51, Oxbow Books, Oxford (2010).
11. Llewellyn-Jones, L.: Aphrodite's Tortoise: The Veiled Woman of Ancient Greece. Swansea: Classical Press of Wales. (2003).
12. Kim, W.-Y.: On the standardization of ceramic terminology. Current Anthropology 11 (2) 168. (1970). https://www.journals.uchicago.edu/doi/abs/10.1086/201121?mobileUi=0
13. Kirkland Lothrop S. Pottery of Costa Rica and Nicaragua, Museum of the American Indian, Heye Foundation, New York. (1926).
14. Claerhout A.: On the standardization of ceramic terminology. Current Anthropology 11(2), 168. (1970).
15. Sarri, K.: Conceptualizing Greek Textile Terminologies: A Databased system. In S. Gaspa, C. Michel, & M.-L. Nosch (Eds.) Textile Terminologies from the Orient to the Mediterranean and Europe, 1000 BC to 1000 AD. pp. 520-527. (2017). Zea Books, Lincoln, NE doi:10.13014/K27P8WJ3
16. ISO 1087-1:2000 Terminology work – Vocabulary – Part 1: Theory and application.
17. Papadopoulou, M.: Headdress for success. Cultic uses of the Hellenistic mitra. In: Brøns C., Nosch, M. L. (eds.) Textiles and Cult in the Ancient Mediterranean, pp. 65-74. Oxbow, Oxford (2017).
18. Sismondo Ridgway, B.: The fashion of the Elgin kore, The Getty Museum Journal 12, 29-58 (1984).
19. Balfet, H., Broutin, Y., Delaporte, Y.: Un essai de système descriptif du vêtement. Vêtement et sociétés 2, L'Ethnographie, vol. 92-94, Société d'ethnographie de Paris, Paris, pp. 363-373 (1984).

20. Delaporte, Y.: Pour une anthropologie du vêtement. Vêtement et sociétés 1, Actes des Journées de rencontre des 2 et 3 mars 1979, Laboratoire d'ethnologie du musée national d'histoire naturelle, Société des amis du Musée de l'Homme, Paris, pp. 3-13 (1981).
21. Eicher, J. B., Roach-Higgins, M. E.: Definition and classification of dress: Implications for analysis of gender roles. In: Barnes R, Eicher J. B. (eds.) Dress and gender: Making and meaning, pp. 8-28. Berg Publishers, New York (NY) (1992).
22. ICOM-Costume Vocabulary of Basic Terms for Cataloguing Costume 1982 http://terminology.collectionstrust.org.uk/ICOM-costume/, last accessed 2016/01/07.
23. The Getty Research Institute, http://www.getty.edu/research/tools/vocabularies/aat/ /, last accessed 2016/01/07.
24. Aimé, X., George, S., Hornung, J.: VETIVOC, une ressource termino-ontologique modulaire du domaine du textile, de la mode et de l'habillement. Revue d'Intelligence Artificielle, vol. 6, pp. 689–728 (2014).
25. ISO 21127:2014-CIDOC-CRM Information and documentation - A reference ontology for the interchange of cultural heritage information. http://www.cidoc-crm.org Definition of the CIDOC Conceptual Reference Model, Version 6.2.3 (May 2018).
26. ISO 704:2009 Terminology work – Principles and methods.
27. Papadopoulou, M. and Roche C.: Ontoterminology of Ancient Greek Garments. In: Roche C. (ed.) TOTh Conference Proceedings, pp. 73-92. (2017).
28. Papadopoulou, M. an Roche C.: Ontologization of Terminology. A worked example from the domain of ancient Greek dress. AIDAinformazioni Journal, 1-2/2018, XXXVI, 89-107. (2018).
29. Humbley, J.: Is terminology specialised lexicography? The experience of French-speaking countries. *Hermes, Journal of Linguistics* 18, 13-31. (1997).
30. Ogden C. K. & Richards I. A.: The meaning of meaning. A study of the influence of language upon thought and the study of symbolism. London: Routledge & Kegan Paul. (1923).
31. Kudashev I. & I. Kudasheva I.: Semiotic Triangle Revisited for the Purposes of Ontology-based Terminology Management. In: C. Roche (ed.). TOTh Conference Proceedings, pp. 83-100. (2010).
32. Roche, C. Le terme et le concept: fondements d'une ontoterminologie. In: Roche, C. (ed.), TOTh Conference Proceedings, pp. 1-22 (2007).
33. Guarino, N., Oberle, D., Staab, S.: What Is an Ontology? In: Staab, S., Studer, R. (ed.) Handbook on Ontologies, Springer-Verlag, Berlin, ISBN 978-3-540-92673-3, pp. 1-16. (2009).
34. Busse, J. et al.: Actually, What Does "Ontology" Mean? A Term Coined by Philosophy in the Light of Different Scientific Disciplines. Journal of Computing and Information Technology - CIT 23, (1), 29–41. (2015).
35. https://www.w3.org/2004/Talks/0412-RDF-functions/slide4-0.html
36. https://www.w3.org/OWL/
37. Musen, M.A.: The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015 (2015). DOI:10.1145/2557001.25757003.
38. Tudorache, T., Nyulas, C., Noy, N. F., & Musen, M. A.: WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web. Semantic web, 4(1), 89-99. (2013).
39. Lavanti, L. Evaluating and designing an ontology editor for domain experts (2018). http://urn.fi/URN:NBN:fi:aalto-201801261168
40. Horridge M., Tudorache T., Vendetti J., Nyulas C., Musen M., Noy N.: Simplified OWL ontology editing for the web: is WebProtégé enough? The Semantic Web - ISWC 2013,

Proceedings part I - 12th International Semantic Web Conference, pp. 200-215. Sydney New South Wales, Australia. (2013).
41. Dzbor, M., E. Motta, C. Buil, J. Gomez, O. Goerlitz, Olaf and H. Lewen Developing ontologies in OWL: An observational study. In: OWL: Experiences and Directions 2006, 10-11-2006, Athens, Georgia, USA. (2006).
42. Suraweera et al.: Using Ontologies to Author Constrained-Based Intelligent Tutoring Systems. In Ducheva D., R. Mizoguchi, J. Greer et al. (Eds). Semantic Web Technologies for E-Learning, (pp. 24-43.) IOS Press, Amsterdam-Berlin, Tokyo, Washington D.C. (2009).
43. LSJ s.v. chiton. In: Perseus 4.0, http://www.perseus.tufts.edu/hopper/resolveform?redirect=true, last accessed 8/1/2019.
44. LSJ s.v. exomis. In: Perseus 4.0, http://www.perseus.tufts.edu/hopper/resolveform?redirect=truelast accessed 8/1/2019.
45. Tedi (ontoTerminology editor), http://new.condillac.org/projects/tedi/, last accessed 8/1/2019.
46. Losfeld, G.: Essai sur le costume grec. Avec 8 planches de l'auteur. Préface de François Chamoux. Paris: Éditions de Boccard. (1991).
47. Desprès S., Roche C. and Papadopoulou, M.: Etude comparative de 2 méthodes outillées pour la construction de terminologies et d'ontologies. Submitted to TOTh 2019. http://toth.condillac.org/, last accessed 8/1/2019.
48. Papadopoulou, M. and Roche, C.: Structuring Humanities' Data through Formal Domain Ontologies: A Use Case from the Domain of Ancient Greek Dress In: EADH 2018: "Data in Digital Humanities", National University of Ireland, Galway 7-9 December 2018Galway, Ireland https://eadh2018.exordo.com/programme/presentation/61
49. Roche C. and Papadopoulou, M.: Définition ontologique du terme. Le cas des vêtements de la Grèce antique. LTT 2018, Lexicologie Terminologie Traduction, Grenoble (France), 27-28 septembre 2018 https://ltt2018.imag.fr/Resumes/Oral/LTT_2018_paper_32.pdf
50. Roche C. and Papadopoulou, M.: Rencontre entre une philologue et un terminologue au pays des ontologies. Revue d'Intelligence Artificielle, Numéro Spécial Humanités Numériques et Intelligence Artificielle (forth.)
51. Papadopoulou, M. and Roche, C.: Tedi: a platform for ontologisation of multilingual terminologies for the Semantic Web SEMAPRO 2018: The Twelfth International Conference on Advances in Semantic Processing. 18-22 November 2018, (2018). Athens, Greece http://www.iaria.org/conferences2018/SEMAPRO18.html
52. Cambridge English Dictionary s.v. collaboration https://dictionary.cambridge.org/dictionary/english/collaboration
53. All European Academies (ALLEA): The European Code of Conduct for Research Integrity 92017) http://www.allea.org/wp-content/uploads/2017/04/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf, last accessed 8/1/2019.
54. Turner, J. R. and Cochrane, R A. The Goals and Methods Matrix: coping with projects with ill-defined goals and/or methods of achieving them. International Journal of Project Management, volume 11, 2 (1993).
55. Dallas, C., et al.: European survey on scholarly practices and digital needs in the arts and humanities. DARIAH, DIMPO. (2017) https://hal.archives-ouvertes.fr/hal-01449002/document, last accessed 8/1/2019.
56. Noy, N., McGuiness, D.: Ontology Development 101: a guide to creating your first ontology / Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, Stanford Medical Informatics Technical Report SMI-2001-0880, Stanford CA. (2001).

Opening up cultural content in non-standard language data through cross-disciplinary collaboration: insights on methods, process and learnings on the example of exploreAT!

Amelie Dorn[1][0000-0002-0848-8149], Yalemisew Abgaz[2][0000-0002-3887-5342] and Eveline Wandl-Vogt[1][0000-0002-0802-0255]

[1] Austrian Centre for Digital Humanities, Austrian Academy of Sciences,Vienna, Austria
[2] ADAPT Centre, Dublin City University, Dublin, Ireland
amelie.dorn@oeaw.ac.at; yalemisew.abgaz@adaptcentre.ie; eveline wandl-vogt@oeaw.ac.at

Abstract. Understanding collaboration between researchers of different disciplines requires an ability to embrace multiple views and perspectives, and communicative efforts. This paper thus provides insights on methods, processes and results of a cooperation in Humanities research supported by semantic technologies with the aim of accessing and opening up cultural knowledge contained in a non-standard language resource. The collaborative undertaking is carried out within a Digital Humanities project and an Open Innovation framework. Meta-disciplinary learnings offer insights on factors fostering mutual understanding, knowledge translation and mutual benefits.

Keywords: Cross-disciplinary cooperation, cross-organizational collaboration, Digital Humanities.

1 Introduction & Background

Culture is a complex phenomenon that offers grounds for analysis in academia, society, arts, etc. from various perspectives [1]. It encompasses several aspects of a society and has been widely expressed and conveyed over the centuries by words, stories, songs, poems, paintings, writings and several other methods most typically through the medium of language. Culture and language are thus tightly interwoven concepts that transcend several societies in time. In recent times, there has thus been a trend in the Humanities in preserving cultural content, mostly contained in written texts, taking language as a first access point. With the support of modern technological tools and the ever growing capacities of digital methods and devices, also otherwise hidden or implicit cultural knowledge contained in Humanities data can be made visible and accessible. Language data thus needs to be available digitally and in technologically enhanced and systematic formats to be accessed and used by the wider community of the modern era, for it to be ultimately preserved through re-use and connectability.

In this paper, we address the collaboration between Humanities scholars and semantic technologists in a Digital Humanities context (the exploreAT! project) [2] on the example of a historic language resource (DBÖ) [3, 4]. We discuss the opening up and exposition of this traditional non-standard German language collection using semantic modelling which exploits existing semantic web standards to represent and facilitate a common representation and interpretation of these cultural resources. Ontologies from different domains and developed by our team are integrated and used to represent the traditional resources to enhance their discoverability and usability in both independent manner and integrated with other similar standardized resource.

We here report on our collaboration results, the humanities background to our research question, the technical methods and implementation, but also on another important yet often unmentioned aspect of language in such cross-disciplinary collaborations, namely the translation of knowledge and expertise across disciplines. Openness to learnings, mutual understanding and communication are key elements in a foundation of bringing about successful results

2 Opening up cultural contents of a traditional language resource: the exploreAT! project

exploreAT! is a current DH project which aims to unveil cultural information contained in a non-standard language resource (DBÖ) [Database of Bavarian dialects in Austria; [3]] by drawing on and combining digital methods and tools from different disciplines (semantic technologies, visualisation prototyping, crowd science) (cf. [5]). At the heart of the project lies the fundamental research question originating from the Humanities background, which asks how to enable access to a non-standard language resource through a cultural lens, giving insights on the conceptualisation of the world and the local society at the time. In this context, the DBÖ resource offers a wealth of not only valuable language data, but also rich cultural content. The database counts around a total of 3.5 million entries, including original data collection questionnaires, answers as well as other digitized excerpts of folklore literature. Originally collected in the area of the former Austro-Hungarian empire with the aim of capturing the speech of the local population, the former collection and following digital preparation was already a huge collaborative effort across persons of various professions, backgrounds and functions, offering detailed documented cultural and societal insights on topics of everyday life (e.g., festivities, professions, nature, food, etc). In particular, our current efforts concentrate around the topic of food, which offers rich grounds for analyses, connectability as well as scientific and societal relevance. Through the support and application of semantic tools, this implicit cultural knowledge can be accessed and connected to other sources and resources for multilingual and multicultural comparison.

3 Cross-cultural Team Communication and Knowledge Exchange: Methods & Tools

The exploreAT! project is all the more interesting as it not only combines cross-disciplinary expertise, but also collaborators of very different cultural and linguistic backgrounds, located across Austria, Spain and Ireland. Methods and tools used for team communication and knowledge exchange are thus key in harmonising and leveraging results and communicating tasks, but also addressing challenges or uncertainties in the workflow. In the wider context of exploreAT!, a combination of digital and analogue methods and tools are employed for ideation (e.g. agile and design thinking tool kits), communication across team members (e.g. web-based project management and communication technologies) or for capturing project ideas and development.
In this paper we concentrate on the description of the specific collaboration scenario which focuses on the creation of the semantic data model. This collaboration arises out of the humanities research question on how to make cultural knowledge in a language resource accessible, discoverable and connectable. In this particular context, current digital tools for communication and task management (Slack, Trello, Skype) were employed, as well as regular face-to-face meetings. While online tools were used for frequent exchange, face-to-face meetings served more specifically for discussions on major project goals, creating work plans or for joint team meetings including also project members. In order to implement collaborative writing, editing or brainstorming a free web-based software office suite was used that could be accessed from any computer with an internet connection.
Drawing on these tools, in what follows we elaborate on the methods, collaborative processes and learnings on the example of the composition of the semantic model [6,7] based on the Humanities research question and resource.

4 Cross-disciplinary Collaboration: the example of creating a Cultural Semantic Data Model

4.1 First processes towards joint collaboration for Semantic Modelling

The aim of the semantic modelling in the context of exploreAT! was to enable the discovery of cultural content in our language collection and connect it to other multilingual and multicultural resources using LOD [8]. The data collection questionnaires and related questions served as the initial access point to the remainder of the collection and to enable connectability to other resources. The modelling further served to understand the semantics of the core entities as defined by the humanists and as contained in the collection, and to represent them and their relationships using existing up-to-date semantic web technologies and standards. With the language collection being focused on a specific domain (non-standard language), and the overall method used to collect the data dating back to the beginning of the 20th century, it was crucial for the semantic

technologists to collaborate in direct exchange with the humanists. In our case, the collaboration involved three major teams. The first team (humanists) consisted of the domain knowledge experts who were involved in or had in-depth knowledge about all steps of the original data collection, organisation and utilisation. The second team (linguists, lexicographers) are researchers in the area of socio-cultural linguistics and related fields, and the third team (technical experts) comprised ontology engineers and semantic web experts responsible for developing the semantic model and uplifting the collection using a linked open data (LOD) platform. The collaboration example we report on here, evolved in three steps.

1. The first joint work laid the foundation for understanding the overall area of expertise and the fundamentals of the dataset.
2. The next step involved collaborating for modelling the core entities of the collection using current semantic web technologies.
3. Finally, search, visualisation and exploitation of the results is presented.

Each of the three steps is described in the following sections.

4.2 Methods and interactions enabling access to implicit data knowledge

Understanding and identifying the implicit knowledge contained in the language collection in general and the detailed meaning and interpretation of the cultural and linguistic entities, in particular, was among the challenges largely faced by both technical experts and linguists. As soon as the semantic modelling process started, the gap became visible in that much of the knowledge which is useful to understand the collection is not self-contained in the data. Thus, it became necessary to gain a deeper knowledge of the data from sources other than the collection itself. Especially for the technical experts, this became a challenge as their objective was to semantically organise and describe the content. Initially, all available information was shared among the teams on the cloud platforms used in the project. This included several resources such as publications describing the collection, notes and change logs. Although the information helped the technical experts to better understand the collection, it generated new questions to the humanists, given the complex structuring of the materials, resulting in less productive weekly meetings and only partially satisfactory advancement. The process became time-consuming as technical experts were remotely located from the humanists, and knowledge experts could not provide the necessary information at the same pace as the technological advancement proceeded.

As communication by digital means only didn't prove optimal, resorting to a different form of knowledge exchange, namely face-to-face meetings, became inevitable. The first face-to-face meeting on the topic of semantic modelling brought the different members involved (humanists, domain experts and technical experts) together in a workshop setting with the aim of building a common understanding of the collection, the methods, resources and techniques used for the original data collection process and to investigate other possible sources of information. This collaboration workshop took

place at exploration space @ ACDH-OeAW in Vienna. The workshop provided valuable insights for both humanists and technical experts as it initiated discussions on topics, such as the identification of cultural content indicators, identification of relevant data fields, or task distribution and enabled the humanists to create new structures for cultural content discovery supporting and enhancing the semantic modelling process. The workshop paved the way for opportunities on planning and proposing a concrete way forward in terms of tasks and workflows, and gave team members a solid understanding of the challenges and complexities involved, and made its contribution to elicit the requirements of each team. Since the initial meeting, a number of similar workshops were conducted in Dublin, Salamanca, Vienna and CERN by incorporating different stakeholders to discuss new opportunities.

Oftentimes a unilateral attempt to model a non-standard language resource can result in an ill-representation, potentially leading to less usability. This face-to-face interaction enabled the discovery of key aspects which would have been challenging, time-consuming or even more complex to communicated by digital or written means only. The semantic modelling exercise resulted in the identification of cultural and linguistic indicators from the side of the humanists and a conceptual model of the collection and its representation using an ontology in owl language, from the technical experts. The resulting ontology and its representation is discussed in detail in [7,9].

A key takeaway for collaboration, is that face-to-face meetings and direct exchange may foster team spirit among collaborators, potentially fuelling further collaboration beyond the current project. In addition, it allows for cross disciplinary collaboration of seemingly far apart areas and benefits members in terms of understanding potential complexities involved in other areas of expertise.

4.3 Synthesizing Humanities and technical expertise towards a first prototype

A next step in the joint collaboration included establishing individual workflows for each team and working towards first common results, a cultural data model for non-standard data questionnaires [6]. Through weekly exchanges and updates using digital communication channels, advancements from both humanists and technical experts were consolidated. Particularly in the joint creation of a data model, the consolidation of views from a semantic web expert and a Digital Humanities are key, as naming conventions or details of representations may vary significantly. Bringing these differences together and narrowing the gap on the representation is crucial, often triggering further revisions, where trade-offs need to be made.

Finally, a first prototype of the data model was presented and discussed with other members of the exploreAT! project in a second workshop. There opportunities arose for the technical expert to engage other project members in a constructive discussion by demonstrating the solution and the application areas. This further face-to-face meeting enabled the technical expert to perform several refinements of the model, including in cleaning noisy data, and it also paved the way for further discussion of the architecture of the implementation. As a result of the direct interaction, several key decisions

could be taken and implemented by all experts involved. Any follow-up communication could thus be continued in online meetings and standups via Skype and Slack channels in regular intervals.

4.4 Creating exploration paths for mutual understanding: facilitating search, visualisation and exploitation of the results

After collaborating in smaller groups for the purpose of elaborating the data model, the next step involved the consolidation and communication of results to the other project members and areas of expertise, such as visual prototyping.
Translating the queries provided by the humanists into a high level technical query language proved challenging. The purpose of the semantic modelling and annotation of the collection was to enable the users to discover cultural content in a non-standard language collection and explore their semantic relationships discovering new insights and support for their research hypotheses. However, providing the resulting semantic research collection with a query user interface often fails in serving the purpose. To address this gap, the initial queries of the humanists were translated to exploration paths in order to elicit the exact requirements. This process involved navigating through the data collection step-by-step, building navigation paths of one or two steps at a time to include further requirements after identifying an initial pivotal query. The exploration paths laid a foundation for the semantic web and visualisation experts to understand the requirements of the users in their own perspectives and to interpret the queries of the target users. At the same time, it enabled the humanists to understand how the semantic data could be efficiently exploited to support their research questions. This was a significant step in the collaboration to understand how the semantic modelling process enhanced the requirements of the users and to provide additional customisable user interfaces to enable the users to pose their own questions.

5 Insights & Conclusion: metadisciplinary learnings

Our collaboration of Humanities research supported by semantic technologies has brought about valuable insights and learnings regarding the knowledge exchange process in terms of creating scientific results, but also in terms of team composition that can prove helpful for training purposes. From our experience, we can report that embracing team diversity brings wealth in both expertise and perspectives. Bringing together researchers of various roles enables a more complete picture and analysis of various perspectives in terms of addressing a particular research question, ultimately consolidating results. What is a key prerequisite, however, is the individual ability to bringing openness and flexibility to a team, which, if lacking, may pose difficulties to the collaboration process. In addition, fostering mutual understanding for involved disciplines can be assured by taking part in training courses in order to obtain basic knowledge in, for example semantic technologies, which also proved beneficial in terms of communication and translation of knowledge. Finally, experimenting and re-

flecting on novel methods of communication and idea-finding may additionally contribute to bringing together different perspectives and enable better mutual understanding. The team applies and analyzes novel approaches towards collaboration in an Open Innovation [10] framework, for example working together with designers [9] to increase the learning curve and potential mutual benefits.

Based on the learnings from exploreAT!, a virtual and physical space for experimentation and innovation has been funded, namely exploration space, currently a best practice example of the Open Innovation platform of the Austrian government (http://open-innovation.gv.at/portfolio/oeaw-exploration-space/).

Acknowledgements

This research is funded by the Nationalstiftung of the Austrian Academy of Sciences under the funding scheme: Digitales kulturelles Erbe, No. DH2014/22 as part of the exploreAT! project, carried out in a collaboration with the Adapt Centre, DCU.

References

1. Longhurst, B. & Baldwin, E. (Eds) Introducing Cultural Studies. Routledge (2008).
2. Wandl-Vogt, E., Kieslinger, B., O'Connor, A. & Theron, R. exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts. In: DHd2015. Von Daten zu Erkenntnissen. 23. bis 27. Februar 2015, Graz. Book of Abstracts. (2015)
3. [DBÖ] Österreichische Akademie der Wissenschaften. (1993–). Datenbank der bairischen Mundarten in Österreich [Database of Bavarian Dialects in Austria] (DBÖ). Wien. [Processing status: 2018/01]
4. Wandl-Vogt, E. …wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikographie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (mit 10 Abbildungen). In P. Ernst (Ed.), Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20. - 23. September 2006. Wien: Praesens, pp. 93–112 (2008).
5. Dorn, Amelie, Eveline Wandl-Vogt, Yalemisew Abgaz, Alejandro Benito Santos, and Roberto Therón. Unlocking Cultural Conceptualisation in Indigenous Language Resources: Collaborative Computing Methodologies. In: Claudia Soria, Besacier, Laurent, and Pretorius, Laurette. (eds.) Proceedings of the LREC 2018 Workshop "CCURL 2018 – Sustaining Knowledge Diversity in the Digital Age", 12 May 2018, Miyazaki, Japan, pp. 19-22 (2018).
6. Abgaz, Yalemisew, Amelie Dorn, Barbara Piringer, Eveline Wandl-Vogt, and Andy Way. Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers. Information 9: 297-320 (2018). doi:10.3390/info9120297
7. Abgaz, Yalemisew, Amelie Dorn, Barbara Piringer, Eveline Wandl-Vogt, and Andy Way. A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis. In: John P. McCrae, Chiarcos, Christian, Declerck, Thierry, Gracia, Jorge, and Klimek, Bettina. Proceedings of the LREC 2018 Workshop "6th Workshop on Linked Data in Linguistics (LDL-2018)". Miyazaki (2018).
8. De Wilde, Max, and Simon Hengchen. Semantic Enrichment of a Multilingual Archive with Linked Open Data. *Digital Humanities Quarterly 11:* 1938 – 4122, (2017).

9. Goikhman, Alisa, Roberto Therón, and Eveline Wandl-Vogt. Designing collaborations: could design probes contribute to better communication between collaborators?. In: Francisco José Garcia-Peñalvo. TEEM '16. Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality. Salamanca, Spain – November 02 - 04, 2016. New York: ACM. (2016) doi:10.1145/3012430.3012431.
10. Open Innovation Strategy for Austria. Goals, Measures & Methods. *Federal Ministry of Science, Research & Economy (bmwfw) and Federal Ministry of Transport, Innovation and Technology (bmvit).* (2015) http://openinnovation.gv.at/wp-content/uploads/2015/08/OI_Barrierefrei_Englisch.pdf, last accessed 2019/01/08