DELIVERABLE

Project Acronym: DM2E

Grant Agreement number: ICT-PSP-297274

Project Title: Digitised Manuscripts to Europeana

D4.6 – Contest Award Fund

Revision: Final 2.0

Authors:

Lieke Ploeger (Open Knowledge Foundation)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	**P**
C	Confidential, only for members of the consortium and the Commission Services	

Revision history and statement of originality

Revision	Date	Author	Organisation	Description
0.1	28.01.2015	Lieke Ploeger	OKFN	Document created
0.2	05.02.2015	Lieke Ploeger	OKFN	Final input added
0.3	06.02.2015	Violeta Trkulja	UBER	Final revision
1.0	09.02.2015	Vivien Petras	UBER	Final Approval
2.0	16.03.2015	Lieke Ploeger	OKFN	Appendix: Adding final version of the reports of the Open Humanities Awards winners

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

Executive summary

This document provides an overview of the Contest Award Fund organised by work package 4, Community Building and Dissemination, during the course of the DM2E project. Under the title 'Open Humanities Awards', two competition rounds took place between 2013-2014 with the goal of rewarding, encouraging and highlighting innovative work building on the technology developed as part of work packages 1, 2 and 3 as well as encouraging partnerships working between developers and non-technical researchers.

A total of over 70 applications were received in these two rounds: the five winning projects communicated their results through monthly updates via the DM2E blog and a final report. They were also provided opportunities to present their work at relevant Digital Humanities and DM2E events.

After a brief introduction on the goal of the contest awards and the two competition rounds, the current deliverable provides further details on the winning projects and their results. Final reports of each of the projects are included as an Appendix.

1 Introduction

One of the key work package 4 tasks was the organisation of a contest award, which forms a significant community building and dissemination opportunity.

Work package 4 decided that the best way to respond to the contest award's objectives of rewarding, encouraging and highlighting innovative work building on the technology developed as part of work packages 1, 2 and 3, as well as encouraging partnership working between developers and non-technical humanities researchers was to run a competition called the Open Humanities Awards. A dedicated website was set up (http://openhumanitiesawards.org) and the awards were published widely across all major Digital Humanities lists.

The first round of the Open Humanities Awards began in year two of the project and focused on supporting open source innovation based on open humanities data. This was in part a practical decision given that the DM2E tools were not yet in a state of readiness to run a compelling competition with. A total of over 50 applications were received and a prestigious judging panel from the Digital Humanities sphere selected two winning projects:

- Dr Bernhard Haslhofer (University of Vienna) for the project 'Maphub'
- Dr Robyn Adams (Centre for Editing Lives and Letters, University College London) for the project 'Joined Up Early Modern Diplomacy'

The next phase of the Open Humanities Awards was launched in year three. This second round consisted of two tracks: an Open track, following on the success of the first round, inviting all submissions using open data or open content, and a dedicated DM2E track, focused on projects building on the research being done in the DM2E project. In this round, a total of 21 applications was received (2 applications to the DM2E track, and 19 to the Open track). The following winners were selected by the high-profile judging panel of Digital Humanities specialists:

Open track:

- Dr. Rainer Simon (AIT Austrian Institute of Technology), Leif Isaksen & Pau de Soto Cañamares (University of Southampton) and Elton Barker (The Open University) for the project 'SEA CHANGE'
- Dr.-Ing. Michael Piotrowski (Leibniz Institute of European History (IEG)) for the project 'Early Modern European Peace Treaties Online'

DM2E track

- Dr. Maximilian Hadersbeck (Center for Information and Language Processing (CIS), University of Munich (LMU)) for the project finderApp WITTFind.

All winning projects communicated their results through monthly updates via the DM2E blog, the DM2E newsletter and a final report. They were also provided opportunities to present their work at relevant Digital Humanities and DM2E events. The current deliverable provides further details on each of the rounds and the winning projects. Final reports of each of the projects are included as an Appendix.

2 Open Humanities Awards

2.1 Round one (2013)

The first round of the Open Humanities Awards was launched in early 2013. In this round, there was a fund of €15,000 worth of prizes on offer for projects using open content, open data or open source tools to further humanities teaching and research. A dedicated website was set up (http://openhumanitiesawards.org) and the awards were published widely across all major Digital Humanities lists and platforms. The full announcement of the first round can be found in the blog post.[1]

In this first round, a total of 50 applications were received. Using many of the experts on the DM2E Digital Humanities Advisory Board and other distinguished voices within the field, WP4 put together a high-profile judging panel who also promoted the awards. The panel included:

- Professor Stefan Gradmann (KU Leuven)
- Dr Susan Schreibman (Trinity College Dublin)
- Professor Andrew Prescott (Kings College London)
- Professor David Robey (Oxford University)
- Dr Melissa Terras (University College London)
- Dr Nicole Coleman (Stanford University)
- Dr Laurent Romary (INRIA)

The winning projects were announced in May 2013. The winners were Bernhard Haslhofer (University Vienna) for a project called Maphub and Dr Robyn Adams (University College London) for the project 'Joined Up Early Modern Diplomacy'. A full introduction on the winners can be found on http://dm2e.eu/open-humanities-award-winners-announced.

Screenshot of the Open Humanities Awards website, round 1 (2013)[2]

Both projects were executed between May 2013 – March 2014 and published regular updates on their results through the DM2E blog. They also presented the results of their

[1] http://blog.okfn.org/2013/02/13/e15000-of-prizes-on-offer-for-open-humanities-projects/.
[2] http://openhumanitiesawards.org/.

award work at the Web as Literature conference (one of the five events run as part of Task 2.4 in Year 2) where they were able to discuss early collaboration with the DM2E work packages. In addition, Bernhard Haslhofer presented the Maphub project at OKCon, the Open Knowledge conference held in September 2013 in Geneva, Switzerland. Final reports from both projects were received at the end of March 2014.

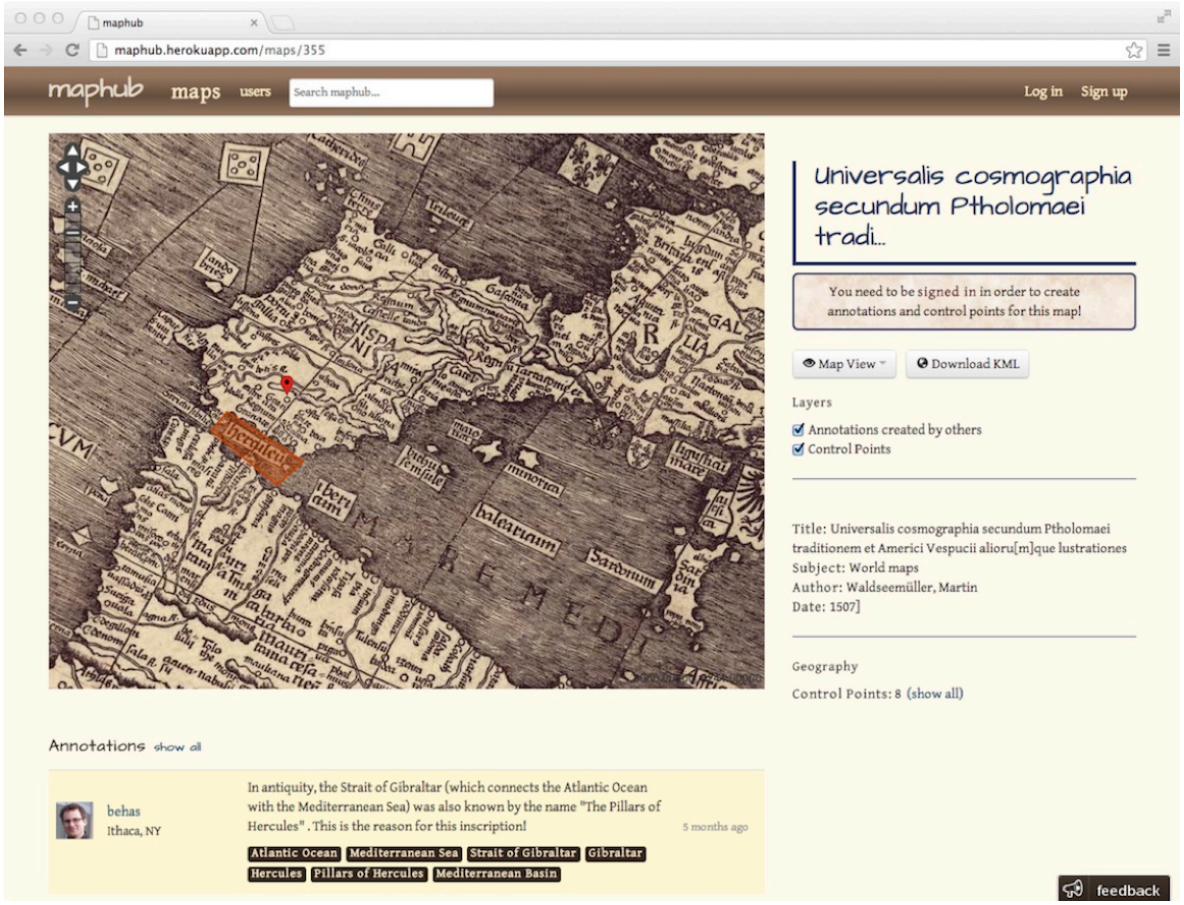
2.1.1 Maphub

The first award went to Dr Bernhard Haslhofer of Vienna University. His project involved building on an open source web application he has been working on called Maphub. Dr Haslhofer told us a little bit about the inspiration for his project:

> *"People love old maps" is a statement that we heard a lot from curators in libraries. This combined with the assumption that many people also have knowledge to share or stories to tell about historical maps, was our motivation to build Maphub.*

In essence Maphub is an open source Web application that pulls out digitised historical maps from closed environments, adds zooming functionality, and assigns Web URIs so that people can talk about them online. During the award period, Maphub was developed as a webportal for annotation of digitised, high-resolution maps, implementing the following major use cases:

1. Annotating regions on high-resolution map images: the high-resolution zoomable maps presented to Maphub users are, in fact, compound Web resources comprising a set of image tiles and a metadata descriptor file. Users have the possibility to zoom into maps and annotate map regions or complete maps.
2. Georeferencing maps: users can mark places on maps (control points) and link those places to geographical web resources. Using this information, it is possible to establish a correspondence between a map's image coordinates and real-world geographic coordinates. This, in turn, enables creation of visual overlays on-top of modern mapping applications.
3. Semantic Tagging: while a user is creating textual annotations on a map or map region, Maphub automatically proposes resources from the Linked Data Web, which may be semantically related to the annotation and therefore also to the underlying annotated map. Users can accept or reject link proposals and thereby create positively or negatively weighted associations between maps and URI-identified Web resources.
4. Sharing Map Annotations: all annotations created in Maphub follow the Open Annotation Data Model specification and are published on the Web as first-class, URI-identified resources. Clients can easily consume map annotations by dereferencing HTTP URIs.



Screenshot of the prototype application of Maphub

Regular blog updates were provided on the ongoing research within the Maphub project:

- Update 1 (July 2013)[3]
- Update 2 (August 2013)[4]
- Final update (April 2014)[5]

In addition, Bernhard Haslhofer presented his findings at the Web as Literature conference in June 2014 and at the Open Knowledge conference (OKCon) in September[6].
The final report[7] on the Maphub project was completed in March 2014 and added to the DM2E website together with the final blog.

2.1.2 Joined Up Early Modern Diplomacy

The second award of round one of the Open Humanities Awards was given to Dr Robyn Adams of Centre for Editing Lives and Letters, University College London. This project proposed to analyse the dataset generated by The Diplomatic Correspondence of Thomas Bodley, 1585-97 by producing visualisations of people and geographical locations mentioned

[3] http://dm2e.eu/open-humanities-awards-maphup-update-1/.
[4] http://dm2e.eu/open-humanities-awards-maphub-update-2/.
[5] http://dm2e.eu/open-humanities-awards-maphub-final-update/.
[6] See slided: http://www.slideshare.net/bhaslhofer/the-story-behind-maphub.
[7] http://dm2e.eu/files/OHAward-Maphub-FinalReport.pdf.

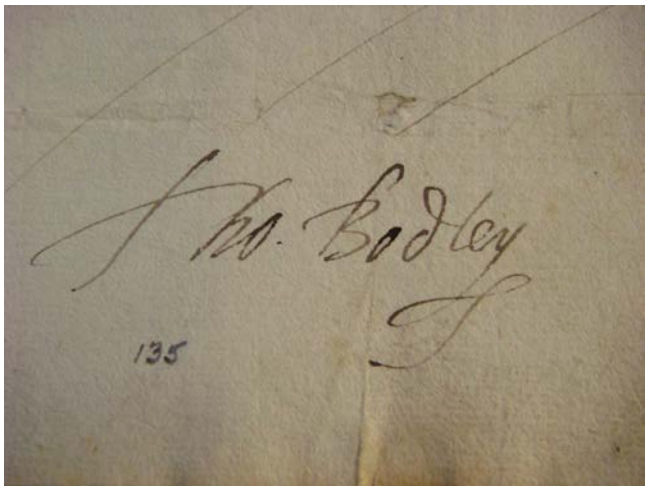
in the letters for a new project, Joined-Up Early Modern Diplomacy: Linked Data from the Correspondence of Thomas Bodley.

The project used 'additional' information that was encoded into the digitisation of early modern letters that took place at the Centre for Editing Lives and Letters. In the initial incarnation of the project this data, which included biographical and geographical information contained within letters was not used (although it was encoded).

The project interrogated three data fields within the larger data set of Bodley's diplomatic correspondence in order to generate visualisations; the network of correspondents and recipients, and the people and places mentioned within the letters. These visualisations were incorporated into the project website to enhance and extend the knowledge derived from the existing corpus of correspondence. The visualisations offer an alternative pathway for scholars and the interested public to understand that in this period especially, the political, university and kinship networks were fundamental to advancement and prosperity.

Following the presentation of their project at the Web as Literature conference in June 2014, the project team of the Centre for Editing Lives and Letters produced regular blog updates on the DM2E blog, as well as a final report [8]in March 2014:

- Update 1 (July 2013)[9]
- Update 2 (August 2013)[10]
- Update 3 (October 2013)[11]
- Update 4 (November 2013)[12]
- Update 5 (January 2014)[13]
- Final update - with final report (April 2014)[14]



[8] http://dm2e.eu/files/JUEMD_report_140422_a.pdf.
[9] http://dm2e.eu/open-humanities-awards-the-diplomatic-correspondence-of-thomas-bodley-update-1/.
[10] http://dm2e.eu/open-humanities-awards-joined-up-early-modern-diplomacy-project-update-2/.
[11] http://dm2e.eu/open-humanities-awards-joined-up-early-modern-diplomacy-update-3/.
[12] http://dm2e.eu/open-humanities-awards-joined-up-early-modern-diplomacy-update-4/.
[13] http://dm2e.eu/open-humanities-awards-joined-up-early-modern-diplomacy-update-5/.
[14] http://dm2e.eu/open-humanities-awards-joined-up-early-modern-diplomacy-final-update/.

2.2 Round two (2014)

After a successful first round, round two of the Open Humanities Awards was launched in April 2014. Since the DM2E tools developed in work packages 1, 2 and 3 were now in a further state of development, this second round consisted of two tracks with a total fund of €20,000 worth of prizes on offer: an Open track (following on the success of the first round) inviting all submissions using open data or open content, and a dedicated DM2E track, focused on projects building on the research being done in the DM2E project. The awards were announced through the Open Humanities Awards website http://openhumanitiesawards.org/, the DM2E blog, the OpenGLAM blog and the Open Knowledge blog, as well as published widely across all major Digital Humanities lists and platforms. The full announcement of the second round can be found on http://dm2e.eu/open-humanities-awards-second-round/.

In the second round, a total of 21 applications was received (2 applications to the DM2E track, and 19 to the Open track). Similar to round one, a high-profile judging panel of Digital Humanities specialists was formed to judge the submissions and further promote the awards. The panel included:

- Professor Andrew Prescott (Kings College London)
- Professor David Robey (Oxford University)
- Dr Melissa Terras (University College London)
- Dr Nicole Coleman (Stanford University)
- Dr Laurent Romary (INRIA)
- Sally Chambers (DARIAH-EU)

The following winners were selected and announced in a blogpost in early July 2014[15]:

Open track:

- Dr. Rainer Simon (AIT Austrian Institute of Technology), Leif Isaksen & Pau de Soto Cañamares (University of Southampton) and Elton Barker (The Open University) for the project 'SEA CHANGE'
- Dr.-Ing. Michael Piotrowski (Leibniz Institute of European History (IEG)) for the project 'Early Modern European Peace Treaties Online'

DM2E track

- Dr. Maximilian Hadersbeck (Center for Information and Language Processing (CIS), University of Munich (LMU)) for the project finderApp WITTFind.

The projects were executed between August 2014 – January 2015 and published regular updates on their results through the DM2E blog. They also presented the results of their award work at the DM2E final event 'Enabling humanities research in the Linked Open Web' on 11 December 2014 in Navacchio, Italy.

The final reports from the SEA CHANGE project was received in January 2015: the final report of the project 'Early Modern European Peace Treaties Online' is due in February 2015. The findings of the DM2E track winner 'FinderApp WITTFind' have been accepted to be presented as an academic paper at the conference "Digital Humanities im deutschsprachigen Raum – DhD 2015"[16] (23-27 February 2015, Graz, Austria).

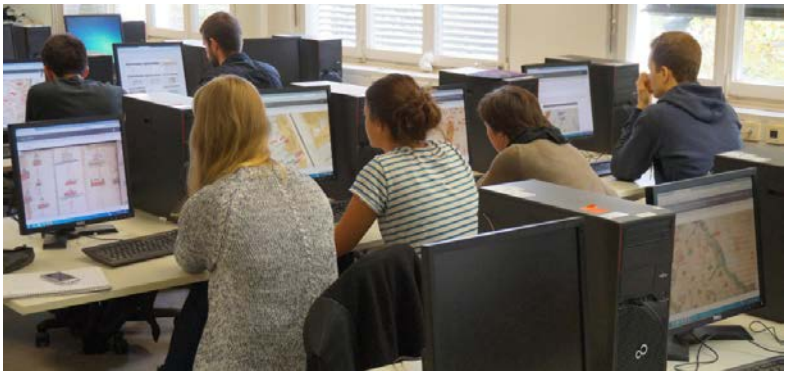
[15] http://dm2e.eu/open-humanities-awards-round-2-winners-announced/.
[16] http://dhd2015.uni-graz.at/.

2.2.1 SEA CHANGE

The first award of the Open track went to Dr. Rainer Simon (AIT), Leif Isaksen & Pau de Soto Cañamares (University of Southampton) and Elton Barker (The Open University) for the project Socially Enhanced Annotation for Cartographic History And Narrative GEography (SEA CHANGE). It complements the ongoing Pelagios research project, a pioneering multi-year initiative funded by the Andrew W. Mellon Foundation, JISC and the AHRC, that aims to aggregate a large corpus of geographic metadata for geospatial documents from Latin, Greek, European medieval and maritime, as well as early Islamic and Chinese traditions.

The project organised two "hackathon"-like workshops, where a mixed audience of students and academics of different backgrounds used Recogito (a web-based tool for the structured annotation of place references in texts and images) to annotate literary texts from the Classical Latin and European Medieval period, as well as Medieval Mappae Mundi and Late Medieval maritime charts.

During these events, participants added an impressive amount of over 15.000 contributions, all of which are now openly available for download and further re-use. The resulting data can be used, for example, to "map" and compare the narrative of the texts, and the contents of the maps with modern day tools like Web maps and GIS; or to contrast documents' geographic properties, toponymy and spatial relationships.



Participants hard at work (1st SEA CHANGE workshop, 31 Oct 2014, Heidelberg)

Contributing to the wider ecosystem of the "Graph of Humanities Data" that is gathering pace in the Digital Humanities (linking data about people, places, events, canonical references, etc.), the project argues that initiatives such as this have the potential to open up new avenues for computational and quantitative research in a variety of fields including History, Geography, Archaeology, Classics, Genealogy and Modern Languages. Most importantly, however, Dr Rainer Simon highlighted at the start:

> *we are convinced that SEA CHANGE is more than just a means to generate exciting new data relevant to humanities research – it is also a chance to engage with a wider audience and, ultimately, build community.*

Regular blog updates were provided on the SEA CHANGE workshops:

- Update 1: First workshop announcement (September 2014)[17]
- Update 2: First workshop report & announcement of second workshop (November 2014)[18]
- Final update: Report of second workshop and final report (December 2014)[19]

[17] http://dm2e.eu/open-humanities-awards-sea-change-update-1/.
[18] http://dm2e.eu/open-humanities-awards-sea-change-update-2/.
[19] http://dm2e.eu/open-humanities-awards-sea-change-final-update/.

In addition, Rainer Simon presented the workshop results at the DM2E final event 'Enabling humanities research in the Linked Open Web' on 11 December 2014 in Navacchio, Italy[20]. The final report on SEA CHANGE was completed in December 2014 and added to the DM2E website together with the final blog.[21]

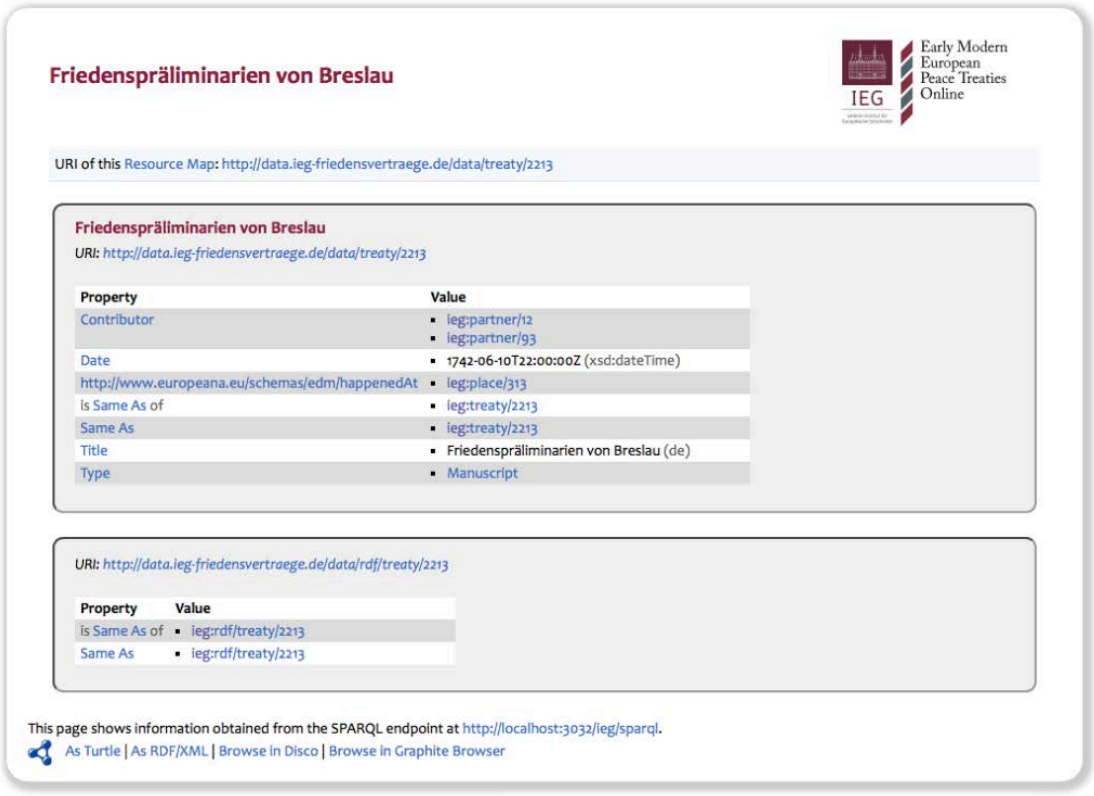
2.2.2 Early Modern European Peace Treaties Online

The second winner in the Open track was Dr.-Ing. Michael Piotrowski (Leibniz Institute of European History (IEG)) for the project Early Modern European Peace Treaties Online ("Europäische Friedensverträge der Vormoderne online"). This is a comprehensive collection of about 1,800 bilateral and multilateral European peace treaties from the period of 1450 to 1789, published as an open access resource by the Leibniz Institute of European History (IEG). The goal of the project funded by the Open Humanities Award is to publish the treaties metadata as Linked Open Data, and to evaluate the use of nanopublications as a representation format for humanities data.

Peace treaties between dynasties and states form an important part of our European cultural heritage. They are also essential for research into early modern peacekeeping and diplomacy. Early Modern European Peace Treaties Online bundles manuscripts that are scattered over archives all over Europe, often hard to access, and partly undocumented. The digitised manuscripts are annotated with basic metadata, and some particularly important treaties are also available as full-text critical editions. This unique combination of digital facsimiles and critical editions has turned out to work as a well-received starting point for scholarly research in this area.

This project brought this valuable collection to the Linked Data cloud, which will allow researchers not only to browse the collection but also to use and reuse the data in novel ways and to integrate it with other collections, including Europeana. It also aimed to represent the key facts of the peace treaties (date, place, signatories, powers, type of treaty, etc.) in RDF using the nanopublications approach, an approach originally developed in the biomedical domain. In addition, there was a link with the DM2E research, as the project used the DM2E model as the basis for their data model.

[20] See slides: http://www.slideshare.net/aboutgeo/sea-change-dm2efinal-conference-pisa-dec-11?ref=http://dm2e.eu/final-dm2e-all-wp-meeting-11-12-december-pisa/.
[21] http://dm2e.eu/files/Linking-Early-Geospatial-Documents.pdf.



Screenshot showing how the information on the Friedenspräliminarien von Breslau (Treaty of Breslau) is presented in Pubby

Regular blog updates were provided on the ongoing research through the DM2E blog:

- Update 1 (September 2014)[22]
- Update 2 (November 2014)[23]
- Update 3 (January 2015)[24]
- Final update (January 2015)[25]

In addition, Michael Piotrowski was present to speak about the results at the DM2E final event 'Enabling humanities research in the Linked Open Web' on 11 December 2014 in Navacchio, Italy[26]. The final report is currently being prepared and will be added to the DM2E website by the end of February.

2.2.3 FinderApp WITTFind

The final award of the series, in the DM2E track, was given to Dr. Maximilian Hadersbeck (Center for Information and Language Processing (CIS), University of Munich (LMU)) for the project finderApp WITTFind. In this project, the research group "Wittgenstein in Co-Text" worked on extending the FinderApp WiTTFind tool to the full Wittgenstein's Nachlass that is made freely available by the Wittgenstein Archives at the University of Bergen and used as linked data software from the DM2E project.

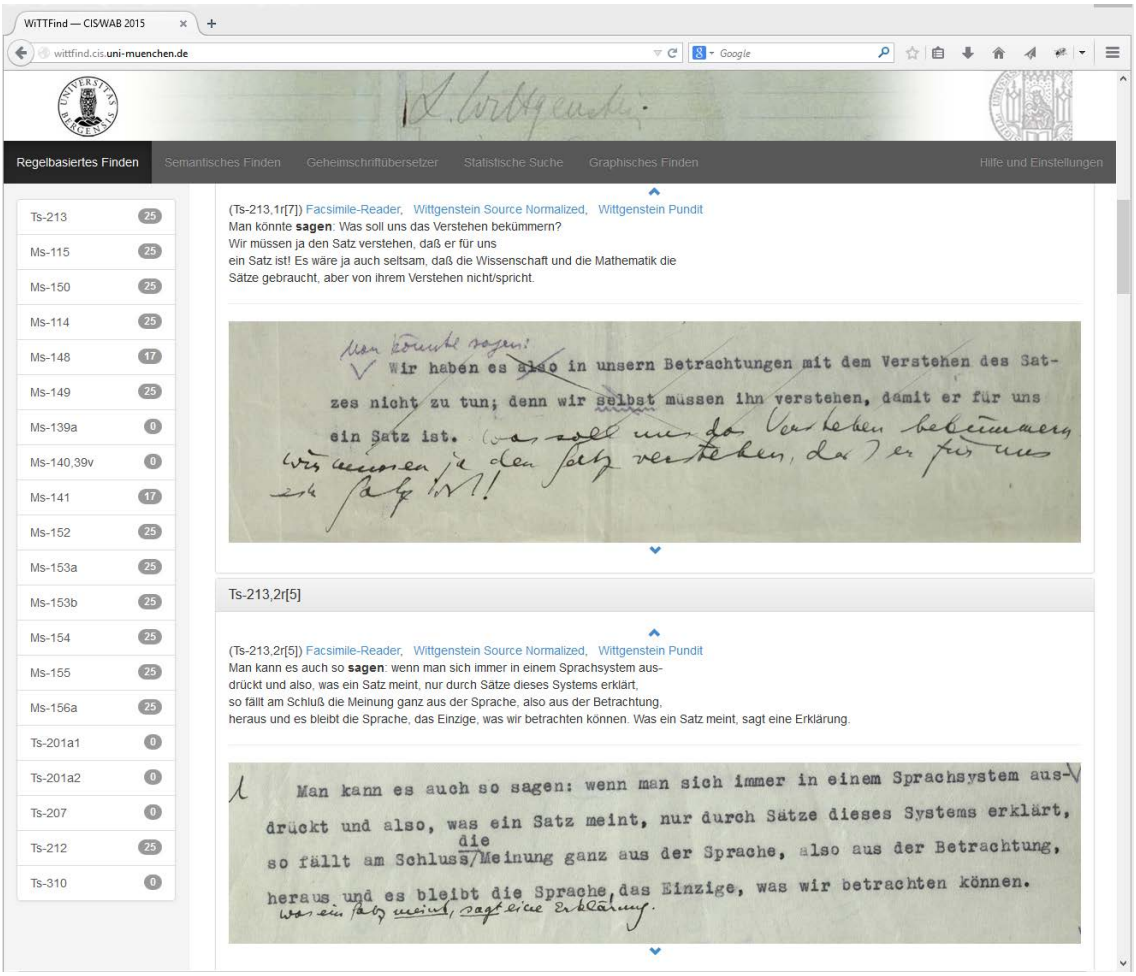
[22] http://dm2e.eu/open-humanities-awards-early-modern-european%20-peace-treaties-online-update-1/.
[23] http://dm2e.eu/open-humanities-awards-early-modern-european%20-peace-treaties-online-update-2.
[24] http://dm2e.eu/open-humanities-awards-early-modern-european-peace-treaties-online-update-3/.
[25] http://dm2e.eu/open-humanities-awards-early-modern-european-peace-treaties-online-final-update/.
[26] See slides: http://www.slideshare.net/DM2E/05-piotrowski.

At his death, the Austrian philosopher Ludwig Wittgenstein (1889-1951) left behind 20,000 pages of philosophical manuscripts and typescripts, the Wittgenstein's Nachlass. In 2009 the Wittgenstein Archives at the University Bergen (WAB), a partner in the DM2E project, made 5,000 pages from the Nachlass freely available on the web at Wittgenstein Source. The research group "Wittgenstein in Co-Text" works on developing the web-frontend FinderApp WiTTFind and the Wittgenstein Advanced Search Tools (WAST), which provide the possibility of rule-based searching the Nachlass in the context of sentences.

The awarded project finderApp WiTTFind offers to the users and researches in the field of humanities a new kind of search machine. Unlike the search capabilities of Google books and the Open Library project, the tools are rule-based and in combination with electronic lexicon and various computational tools, this project provides lemmatised and inverse lemmatised search and allows queries to the Nachlass which include word forms, semantic and sentence structured specifications. Syntactic disambiguation is done with Part-of-Speech tagging. Query results are displayed in a web browser as XSLT-transformations of the transcribed texts, together with facsimile of the matching segment in the original. With this information researchers are able to check the correctness of the edition and can explore original handwritten edition-texts which are otherwise stored in access-restricted archives.

The project consisted of three elements:

- An extension of the finderApp which is currently used for exploring and researching only Ludwig Wittgenstein's Big Typescript TS-213 (BT) to the rest of the openly available 5,000 pages of Wittgenstein's Nachlass
- Making the tool openly available to other humanity projects by defining APIs and a XML-TEI-P5 tagset, which defines the XML-structure of the texts which are processed from the finderApp
- Building a git-server-site which offers the applications and programs to other research projects in the field of Digital Humanities



The new multidoc webpage (http://wittfind.cis.uni-muenchen.de)

Maximilian Hadersbeck provided monthly blog updates on the ongoing research through the DM2E blog:

- Update 1 (September 2014)[27]
- Update 2 (October 2014)[28]
- Update 3 (November 2014)[29]
- Update 4 (January 2015)[30]
- Final update (January 2015)[31]

In addition, the research findings were presented at the DM2E final event 'Enabling humanities research in the Linked Open Web' on 11 December 2014 in Navacchio, Italy[32].

A final aim of the project was to place a publication of the work at an important congress. A paper and a poster were submitted to the "Digital Humanities im deutschsprachigen Raum – DhD 2015"[33] conference (23-27 February 2015, Graz, Austria). Both were evaluated well

[27] http://dm2e.eu/open-humanities-awards-finderapp-wittfind-update-1.
[28] http://dm2e.eu/open-humanities-awards-finderapp-wittfind-update-2.
[29] http://dm2e.eu/open-humanities-awards-finderapp-wittfind-update-3.
[30] http://dm2e.eu/open-humanities-awards-finderapp-wittfind-update-4.
[31] http://dm2e.eu/open-humanities-awards-finderapp-wittfind-final-update/.
[32] http://www.slideshare.net/DM2E/09-pisa-finale.
[33] http://dhd2015.uni-graz.at/.

and accepted. In Graz, Maximilian Hadersbeck will be presenting the paper “Wittgensteins Nachlass: Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind”. Co-authors to the text are Alois Pichler, Florian Fink, Daniel Bruder and Ina Arends. The poster which will be presented gives a demo of the project and has the title “Wittgensteins Nachlass: Aufbau und Demonstration der FinderApp WiTTFind und ihrer Komponenten.” The authors of the poster are Yuliya Kalasouskaya, Matthias Lindinger, Stefan Schweter and Roman Capsamun. Both presentations can be found in the programme of the conference and will be added to the DM2E website following the event.[34]

[34] https://www.conftool.pro/dhd2015/sessions.php.

Appendix: Final reports

Round 1 (2013)

- Maphub
- Joined Up Early Modern Diplomacy

Round 2 (2014)

- SEA CHANGE
- Early Modern European Peace Treaties Online
- FinderApp WITTFind

Maphub - Final Report

1. Introduction

Historic maps record historical geographical information often retained by no other written source (Rumsey and Williams, 2002) and give insight into socio-economic and environmental phenomena such as land use, river channel changes or flood (e.g., Pearson, 2006; Braga, G., & Gervasoni, 1989; Witschas, 2003). They allow the reconstruction of past urban environments (e.g.; Isoda et al., 2010) and draw a picture of the cultural, political and religious context in which they were created (Rumsey and Williams, 2002). Their geographical accuracy tells us much about the state of technology at the time of their creation. Consequently, historic maps are cultural heritage artifacts in their own right, part of the artistic heritage as much as of the history of science and technology as a whole (Boutoura and Livieratos, 2006).

Scholars who study these maps often want to take notes on certain maps or map regions, view certain areas in the context of today's maps, associate map regions with historical events, places, or even persons. Annotations are a fundamental scholarly practice common across disciplines (Unsworth, 2000) and a scholarly primitive that enables scholars to organize, share and exchange knowledge, and work collaboratively in the interpretation and analysis of source material. At the same time, annotations offer additional context: they supplement the item under investigation with information that may better reflect a user's setting (Frisse, 1987). However, many historic maps, which have been digitized so far, still reside in closed system environments within libraries, museums, or private collections (e.g., Rumsey Historical Map Collection[1]). Those that are already published on the Web don't allow scholars or end-users to annotate them in a way that is interoperable across systems.

Therefore we built a demonstrator entitled *Maphub*, which is a Web portal allowing annotation of digitized, high-resolution maps. It implement five major use cases:

1. **Annotating regions on high-resolution map images**: the high-resolution zoomable maps presented to Maphub users are, in fact, compound Web resources comprising a set of image tiles and a metadata descriptor file. Users have the possibility to zoom into maps and annotate map regions or complete maps.
2. **Georeferencing maps**: users can mark places on maps (control points) and link those places to geographical Web resources (e.g., Geonames[2]). Using this information, it is possible to establish a correspondence between a map's image coordinates and real-world geographic coordinates. This, in turn, enables creation of visual overlays on-top of modern mapping applications (e.g., Google Earth).
3. **Semantic Tagging**: while a user is creating textual annotations on a map or map region, Maphub automatically proposes resources from the Linked Data Web (e.g., DBPedia), which may be semantically related to the annotation and therefore also to the underlying annotated map. Users can accept or reject link proposals and thereby

[1] http://www.davidrumsey.com/
[2] http://www.geonames.org/

create positively or negatively weighted associations between maps and URI-identified Web resources.

4. **Sharing Map Annotations**: all annotations created in Maphub follow the Open Annotation Data Model specification and and are published on the Web as first-class, URI-identified resources. Clients can easily consume map annotations by dereferencing HTTP URIs.

For demonstration purposes, we bootstrapped the portal with 6000 digitized historical maps taken from the Library of Congress Historic Map division. Those maps are not covered by copyright protection and can easily be reused without technical, financial, or legal restrictions. In the following, we will elaborate the conceptual and technical details of each use case. We will conclude this report with lessons learned from building the Maphub demonstrator and briefly discuss how selected system components are being further developed and how they can be reused in other applications.

2. Annotating Maps

Maphub is available in any modern Web browser and organized as an open source project[3]. It allows users to retrieve maps either by browsing or searching over available metadata and user-contributed annotations and tags. Users can zoom into maps, highlight a region on the map, and add their knowledge about that region by adding textual annotations. Figure 1 shows the central Maphub map annotation view.

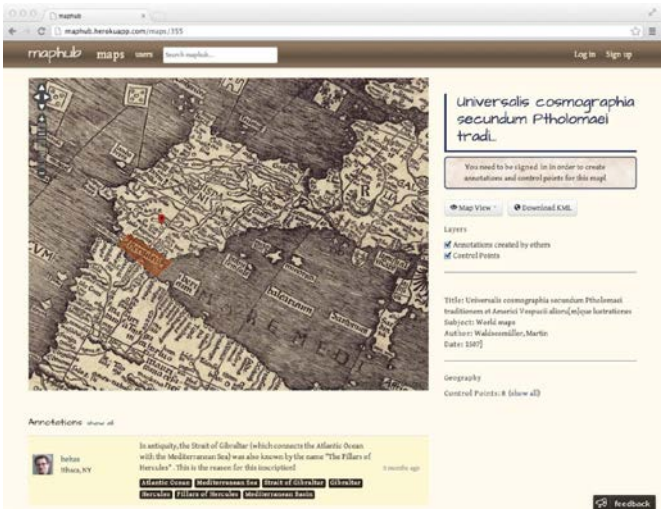


Figure 1. Maphub Map Annotation View.

To create an annotation, users markup regions on the map with geometric shapes such as polygons or rectangles. Once the area to be annotated is defined, they are asked to tell their stories and contribute their knowledge in the form of textual comments. While users are composing their comments, Maphub periodically suggests tags based on either the text contents or the geographic location of the annotated map region. Suggested tags appear below the annotation text. The user may accept tags and deem them as relevant to their annotation or reject non-relevant tags. Unselected tags remain neutral.

[3] http://maphub.github.io

Figure 2 shows an example user annotation created for a region covering the Strait of Gibraltar. While the user entered a free-text comment related to the naming of the area, Maphub queried an instance of Wikipedia Miner[4] to perform named entity recognition on the entered text and received a ranked list of Wikipedia resource URIs (e.g., http://en.wikipedia.org/wiki/Mediterranean_sea) in return. URIs should not be exposed to the user, so Maphub displays the corresponding Wikipedia page titles instead (e.g., Mediterranean Sea). Since page titles alone might not carry enough information for the user to disambiguate concepts, Maphub offers additional context information: the short abstract of the corresponding Wikipedia article is shown when the user hovers over a tag.

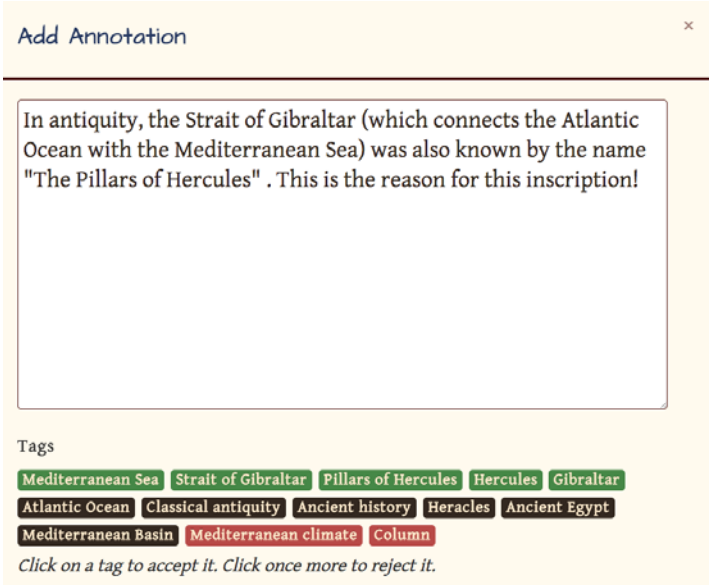


Figure 2. Maphub Annotation Input Dialogue.

Once tags are displayed, users may mark them as relevant for their annotation by clicking on them once, which turns the labels green. Clicking once more rejects the tags, and clicking again sets them back to their (initial) neutral state. In the previous screenshot, the user accepts five tags and actively prunes two tags that are not relevant in the context of this annotation.

3. Georeferencing Maps

Besides commentarial annotations, which have been described in the previous section, Maphub also support so-called Georeference Annotations, which allows users to create an annotation with the intention to express a correspondence between a point/region on the map and either a point/region in a defined geographic coordinate system or an authoritative Gazetteer. The goal of this type of annotation is to establish *control points* for raster image maps. The current Maphub application uses Geonames as Gazetteer and links control points to URI-identified locations, which provide further information such as latitude and longitude coordinates. Figure 3 shows examples of control points added to historic maps.

[4] http://wikipedia-miner.cms.waikato.ac.nz/

Figure 3. Control Point Annotation Examples.

After at least three of these control points are added to a map, a geographical model can be computed for the map. This allows Maphub to prompt the user with more locations and suggest those locations as semantic tags in the annotation input dialogue.

Furthermore, after adding at least three control points to a map, it is possible to calculate real-world locations for any point on the map and create overlay views on modern mapping applications such as Google Maps or Google Earth. These views will overlay a historic map onto its current day location. Figure 4 shows example map overlays.

Figure 4. Google Maps and Earth Overlays created from georeferenced historic map images.

4. Semantic Tagging

Maphub's semantic tagging feature has been motivated by the problem that despite their wide-spread adoption, tagging systems still face a number of problems: a tag can be ambiguous and have many related meanings (*polysemy*), multiple tags can have the same meaning (*synonymy*), or the semantics of a tag might range from very specific to very general because people describe resources along a continuum of specificity (Golder and Huberman, 2006). These issues are rooted in label-based nature of tags and important for system providers who want to exploit the semantics and contextual information associated with tags for resource discovery. If, for instance, a user tags a resource with *Paris* it is not entirely clear whether this tag means *Paris*, the capital of France or *Paris*, the city in the United States. Contextual information, such as the translations of the term *Paris* in other world-languages or its geographical location can only be determined after reconciling label-based tags with data entries in other data sources.

Mapping label-based tags to concepts defined in knowledge contexts, such as Wikipedia is a possible solution. Sigurbjörnsson and Van Zwol (2006) use string matching to map Flickr tags to WordNet semantic categories and found that 51.8% of the tags in Flickr can be mapped. Overell et al. (2009) use concept definitions from Wikipedia and Open Directory to classify tags automatically and show that nearly 70% of Flickr tags can be classified correctly. However, in all these approaches tag semantics is determined heuristically and a-posteriori, without taking into account the user who created and assigned the tag and knows about its precise semantics.

To solve this problem, we propose that users associate URI-identified Web resources from a knowledge context, such as Wikipedia, as part of their tagging activity. A tagging system could suggest the label *Paris* as a possible tag in the user-interface, but create a link to a Web resource (e.g., http://en.wikipedia.org/wiki/Paris) in the back-end. We call this technique *semantic tagging*. Different from label-based tagging, the semantics of a tag is determined by its creator at creation time. Each tag also leads to further contextual information that can be exploited for resource discovery purposes. Explicit user feedback on suggested tags results in a graph of positive and negative tagging relationships that can be used to improve tag recommendation strategies.

To demonstrate the user acceptance of this approach, we implemented semantic tagging in Maphub. We wanted to illustrate how to design semantic tagging systems so that users can easily select from suggested semantic tags, accept or reject proposed tags, without ever having to interact with URIs directly. We also ran an empirical evaluation to compare semantic tagging with other tagging techniques. In the following we discuss the conceptual and design-related aspects of the semantic tagging technique and compare it with existing, label-based tagging design characteristics. We also briefly summarize the main findings of our experiments, which are described in more detail in Haslhofer et al. (2013).

4.1. Conceptual Model

In the conceptual model for label-based tagging systems introduced by Marlow et al. (2006), which is shown in Figure 5, a user *u* assigns a tag *t* to a resource *r*. Tags are represented as labeled edges that connect users and resources but do not carry or refer to any additional contextual information. Both resources and users may be connected to other nodes, since there may be links between Web pages and users may belong to social networks. Label based tagging systems can allow for multiplicity of tags around resources (*bag-model*) or deny tag repetitions (*set-model*).

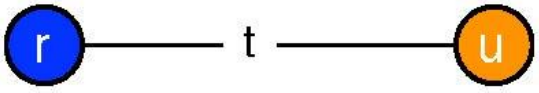


Figure 5. Label-based Tagging Model.

Semantic tagging, which is shown in Figure 6, extends this model by representing a tag *t* as a qualified relationship between two resources: r_x is the resource identifying and defining the semantics of a tag (e.g.,http://en.wikipedia.org/wiki/Paris), and r_y is the resource being tagged (e.g., a photo taken in Paris). The former is defined within a knowledge context *K* and

can carry textual labels (e.g., *Paris*) and additional context information (e.g., Paris is a city in France). Possible knowledge contexts are online encyclopedias such as Wikipedia, place name registries such as GeoNames, structured Web data sources such as Freebase[5], domain-specific Web vocabularies or gazetteers, or any other Linked Data source providing suitable concept definitions. An explicit, qualified semantic tagging relationship also implies an *about* relationship between the involved resources, meaning that r_x is about r_y if they are connected by a user via a semantic tagging relationship.

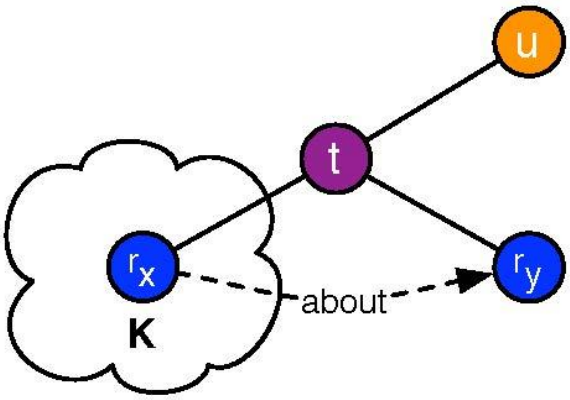


Figure 6. Semantic Tagging Model.

Since semantic tags can also be represented as first-class URI-identified Web resources, the resulting model is not label- or set-based but *graph-based*, with different types of nodes (users, resources) being connected to each other. This enables multiplicity and aggregation of tags not only around resources but also around users and user groups, which can be exploited for graph-based tag recommendation and user-based collaborative tag filtering.

We believe that an information system implementing semantic tagging should allow users to easily select from suggested tags, accept or reject proposed tags, without ever having to interact with URIs. Therefore we will now continue discussing the following design aspects in more detail: *tag recommendation*, *user feedback*, and *user transparency*.

4.2. Tag recommendation

Marlow et al. (2006) distinguish between three main categories existing systems fall into: *blind tagging*, where a user cannot view tags assigned to the same resource by other users, *viewable tagging*, where users sees tags associated with a resource, and *suggestive tagging*, where the system suggests possible tags to the user. Suggestive tagging systems can derive tags from existing tags by the same or other users or gather them from a resource's context.

Following this classification, we perceive semantic tagging as a special form of suggestive tagging, where tag resources are recommended from a given knowledge context, based on the context of any resource that is part of the semantic tagging graph. As in other suggestive tagging systems (see Gupta et al., 2010), tag recommendation strategies can consider the content (e.g., image file) or context (e.g., metadata, other tag resources) of a resource. If the

[5] http://www.freebase.com/

applied knowledge context follows a graph structure, it is also possible to apply graph-based recommendation strategies for tag resource proposals. When, for instance, a system proposes the semantic tag *Paris*, it could also propose related resources such as *France*, and *Eiffel Tower* if these concepts are semantically connected in the underlying knowledge context - as it is the case in Wikipedia. In Maphub, for example, we recommend semantic tags based on the text users are entering while they are authoring annotations on historical maps.

Semantic tag suggestion can be implemented by calling named entity recognition services that link things mentioned in plain text to Web resources, such as Wikipedia Miner[6] or DBPedia Spotlight[7].

4.3. User feedback

Adding a label-based or semantic tag to a given resource usually means that the tag is somehow about or describes the resource, at least within the context of the tag creator. If a user applies the tag *Paris* to an image it is assumed that Paris is somehow about that image. Thus, an intrinsic assumption of existing tagging models is that relationships between tags and resources have positive connotations.

However, with tags becoming first-class resources describing a qualified relationship between resources, one can also capture negative relationships: when the system recommends a set of possibly relevant (semantic) tags and the user accepts one of them, it can infer a positive tagging relationships. However, the system could also capture the non-accepted or explicitly rejected tags and interpret them as negative tagging relationships, as illustrated in Figure 7. An explicitly rejected tag *Berlin* on an image showing Paris is an example for such a negative relationship.

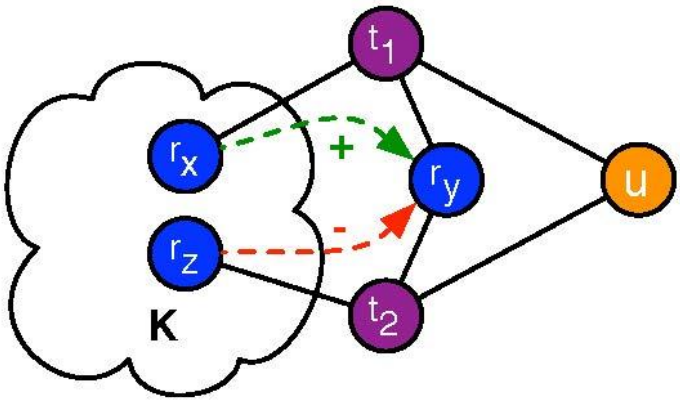


Figure 7. Semantic Tags forming a Graph of Positive and Negative Relationships.

Qualified semantic tagging relationships carrying positive and negative weights can easily be transformed into a bipartite graph of positive (accept) and negative (reject) *about* relationships between semantic tags and tagged resources. From this graph, one can

[6] http://wikipedia-miner.cms.waikato.ac.nz/
[7] https://github.com/dbpedia-spotlight

directly derive relevance judgments for given pairs of Web resources and build gold standards, which are required for subsequent information retrieval tasks.

4.4. User transparency

The World Wide Web uses HTTP URIs to unambiguously identify Web resources, such as the Wikipedia article about Paris. However, URIs are opaque strings that do not necessarily carry any semantics. While the design choice in Wikipedia was to use-human readable URIs (e.g.,http://en.wikipedia.org/wiki/Paris), other sources do not follow this approach. In the GeoNames knowledge context, for instance, Paris is identified by a URI with a numeric path element http://www.geonames.org/2988507. Such a URI syntax is hard to remember for human end-users and might lead to errors when being transcribed manually.

Therefore, semantic tagging systems should hide the technical aspects of this approach underneath the user-interface and follow the design of existing suggestive tagging interfaces: they should neither display nor prompt users to input HTTP URIs, but suggest labels and maintain internal, user-transparent mappings between labels and their corresponding resources. For example, instead of displaying a semantic tag URI for Paris the system should present labels such as *Paris*.

This of course requires that the knowledge context also provide human-readable labels for resource definitions, which is common practice in real-world data sources. In the case of Wikipedia one can, for instance, extract the article's title (*Paris*) directly from the Web page or rely on DBpedia, which provides structured data extracted from Wikipedia.

4.5. Empirical Evaluation Summary

While working on Maphub, its semantic tagging functionality has become our core research interest. We conducted an in-lab user study with 26 participants to find out how semantic tagging differs from label-based tagging and other suggestive techniques. Our central findings can be summarized as follows:

- Our semantic tagging implementation does not affect tag production, the types and categories of obtained tags, and user task load, while providing tagging relationships to well-defined concept definitions.
- When compared to label-based tagging, our technique also gathers positive and negative tagging relationships, which can be useful for improving tag recommendation and resource retrieval.

Hence, semantic tagging as implemented in Maphub could produce the same result as a label-based tagging, with the main difference that semantic tagging gives references to unambiguous Web resources instead of semantically ambiguous labels. More details on the methodology and results of that experiment are described in our report available at (http://arxiv.org/abs/1304.1636).

5. Sharing Map Annotations

Sharing collected annotation data in an interoperable way was another major development goal. Maphub is an early adopter of the Open Annotation model[8] and demonstrates how to apply that model in the context of digitized historic maps and how to expose comments as well as semantic tags. As described in the Maphub API[9] documentation, each annotation becomes a first class Web resource that is dereferenceable by its URI and therefore easily accessible by any Web client. In that way, while users are annotating maps, Maphub not only consumes data from global data networks - it also contributes data back. In the following we briefly introduce the central aspects of the Open Annotation model and describe how we implemented it in Maphub.

5.1. Open Annotation Data Model

Annotations on the Web have many facets: a simple example could be a textual note or a tag annotating an image or video. Things become more complex when a particular paragraph in an HTML document annotates a segment in an online video or when someone draws polygon shapes on tiled high-resolution image sets, such as the historical maps used in Maphub. Therefore in a generic, Web-centric conception, an annotation can be regarded as an association between a body and a target resource (Haslhofer et al., 2011) .

Annotea (Kahan, 2002) already defines a specification for publishing annotations on the Web but has several shortcomings: (i) it was designed for the annotation of Web pages and provides only limited means to address segments in multimedia objects, (ii) if clients want to access annotations they need to be aware of the Annotea-specific protocol, and (iii) Annotea annotations do not take into account that Web resources are very likely to have different states over time.

Throughout the years several Annotea extensions have been developed to deal with these and other shortcomings: Koivunnen (2006) introduced additional types of annotations, such as *bookmark* and *topic*. Schroeter et al. (2007) proposed to express segments in media-objects by using \emph{context} resources in combination with formalized or standardized descriptions to represent the context, such as SVG or complex datatypes taken from the MPEG-7 standard. Based on that work, Haslhofer et al. (2009) introduce the notion of *annotation profiles* as containers for content- and annotation-type specific Annotea extensions and suggested that annotations should be dereferenceable resources on the Web, which follow the Linked Data guidelines. However, these extensions were developed separately from each other and inherit some of the above-mentioned Annotea shortcomings.

In 2011 the Open Annotation Collaboration (OAC)[10] formed as an international group with the aim of providing a Web-centric, interoperable annotation environment that facilitates cross-boundary annotations, allowing multiple servers, clients and overlay services to create, discover and make use of the valuable information contained in annotations. A Linked Data based approach has been adopted and resulted in the formation of the W3C Open

[8] http://www.openannotation.org/spec/core/
[9] http://maphub.github.io/api
[10] http://www.openannotation.org/

Annotation Working Group, which recently published a first Open Annotation Community Draft[11]. Figure 8 shows the core conceptual model of the current model specification.

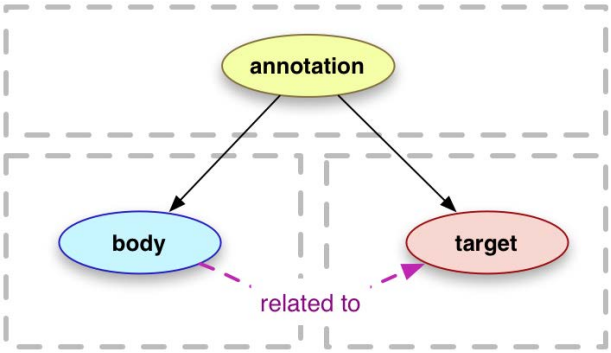


Figure 8. Open Annotations Data Model - Core Model.

Maphub is an early adopter of the Open Annotation model and demonstrates how to apply the model in the context of annotations on historic maps and how to expose georeference and commentarial annotations as well as semantic tags as first class Web-resources that are dereferenceable by their URIs. In that way, while users are annotating maps, Maphub not only consumes data from open data sources - it also contributes open data back. In the following we describe how Maphub implements the Open Annotation Data model for the types of annotations it currently supports.

5.2. Sharing Georeference Annotations

A Georeference Annotation associates a place URI, which can be interpreted as a semantic tag, with a place on the map (the annotation Target). Place URIs are provided by the Geonames online gazetteer (e.g., London, UK: http://sws.geonames.org/2643743/). Georeference Annotations in Maphub are dereferenceable Web resources. When a client issues an HTTP GET request against the Georeference Annotation HTTP URI, Maphub determines the response format based on the value of the HTTP Accept header submitted by the client.

Figure 9 shows an example Georeference Annotation represented in the Open Annotation model. Each annotation receives its own URI (yellow) and follows one more annotation types (e.g., oa:Annotation). Common types can be defined as part of the (extended) Open Annotation specification or introduced on a per-application basis (e.g, maphub:GeoReference). Descriptive metadata can be attached to each annotation (e.g.: annotation author information). In this case, the annotation's body is a semantic tag, i.e., the place identified by a GeoNames URI, whereas the target is a specific resource, which represents the highlighted region on the map, identified by x,y, width, and height parameters.

[11] http://www.openannotation.org/spec/core/

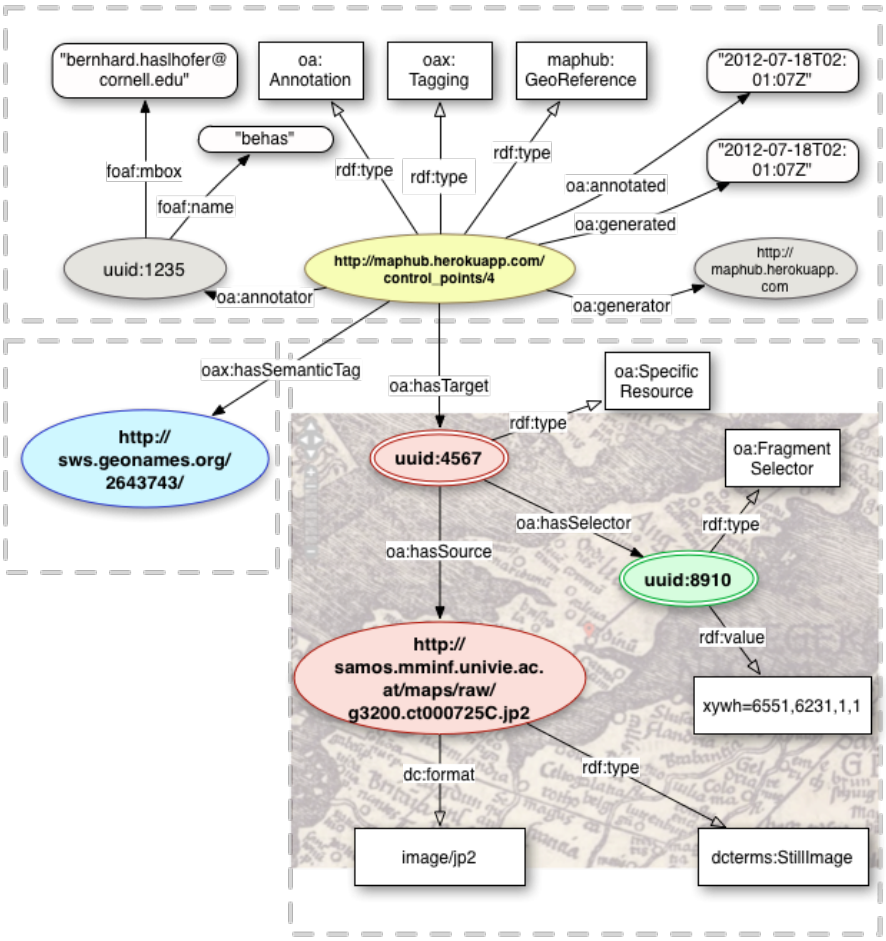


Figure 9. Example Georeference Annotation exposed as Open Annotation.

5.3. Sharing Commentarial Annotations and Semantic Tags

Figure 10 illustrates how this annotation is represented in the Open Collaboration Model. The annotation text is represented as an Inline Body, and the semantic tags as Semantic Tags. Since the annotation is about part of the map resource, the annotation target is a Specific Target, which is further described by two Selector representations: one SVGSelector and a custom Selector that expresses the same information in the Well-known text (WKT) markup language, which is commonly used in geographic information systems.

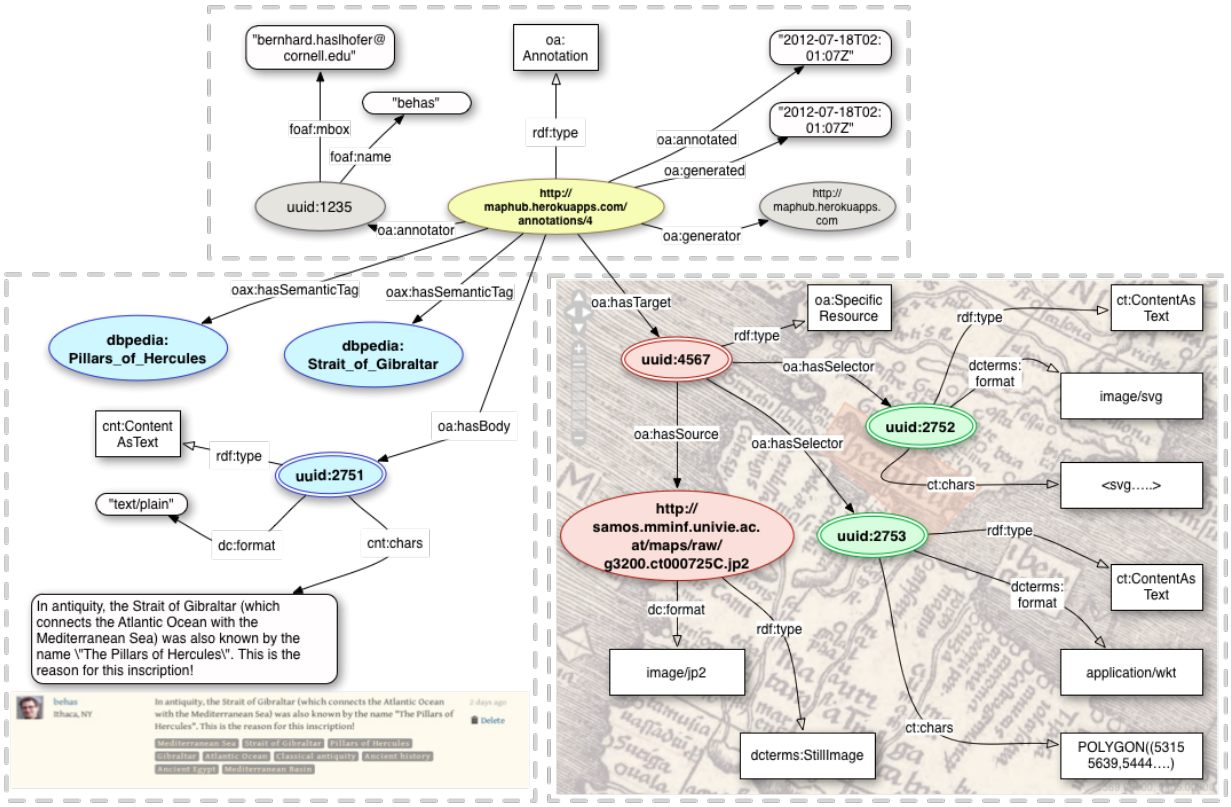


Figure 10. Example Commentarial Annotation exposed as Open Annotation.

Besides exposing individual georeference and commentarial annotations, Maphub also exposes annotation indices that enable discovery of those annotations.

6. Lessons Learned and Future Work

With Maphub we demonstrated how a Web-based approach could support scholars who study historical maps in taking notes on certain maps or map regions, viewing certain areas in the context of today's maps, and associating map regions with historical events, places, or even persons. We believe that semantic tagging is a key feature in such a process and findings from our empirical evaluation confirmed that this feature is worth to be further explored.

Overall, we believe that our findings carry implications for designers who want to adopt semantic tagging in other scenarios. A major incentive for system providers to implement tagging is to obtain metadata describing the content and context of online resources, which is important for efficient resource discovery but expensive in terms of time and effort when created manually. In traditional, label-based tagging systems providers can add possibly ambiguous label- based tags to their records. With semantic tagging, they obtain references to concepts defined in other Web-based knowledge context. Traditional information retrieval techniques can be enhanced to exploit these relationships and consider additional contextual information.

We believe that people might also want to annotate other things on the Web and that Web annotation tools should support semantic tagging as well. Therefore, we will make it

available as a plugin for Annotorious[12], which is a JavaScript image annotation library that can be used in any Website, and is also compatible with the Annotator[13] Web annotation system. Figure 11 shows how semantic tagging can now be applied for any Web image using the Annotorious library.

Figure 11. Semantic Tagging in Annotorious.

Finally, we would like to emphasize that availability of **open metadata and content** - in our case the historic map collection from the Library of Congress - has been key for designing and implementing Maphub and for experimenting with previously unavailable features. Availability of open APIs and absence of copyright restrictions allowed us to bootstrap Maphub with minimal technical and no financial or legal effort.

[12] http://annotorious.github.io
[13] http://okfnlabs.org/annotator

References

C. Boutoura and E. Livieratos. (2006). Some fundamentals for the study of the geometry of early maps by comparative methods. e-Perimetron, 1(1):60–70, 2006.

Braga, G., & Gervasoni, S. (1989). Evolution of the Po River: an Example of the Application of Historic Maps. In H. M. G.E. Petts and A. Roux, editors, Historical Change of Large Alluvial Rivers: Western Europe, pages 113-126, New York, Wiley.

Frisse, M. (1987). Searching for Information in a Hypertext Medical Handbook. In Proceedings of the ACM Conference on Hypertext (HYPERTEXT '87)

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of information science*, *32*(2), 198-208.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, *12*(1), 58-72.

Haslhofer, B., Simon, R., Sanderson, R., & Van de Sompel, H. (2011). The open annotation collaboration (OAC) model. *Multimedia on the Web (MMWeb), 2011 Workshop on*. IEEE.

Haslhofer, B., Jochum, W., King, R., Sadilek, C., & Schellner, K. (2009). The LEMO annotation framework: weaving multimedia annotations with the web. *International Journal on Digital Libraries*, *10*(1), 15-32.

Haslhofer, B., Robitza, W., Guimbretiere, F., & Lagoze, C. (2013). Semantic tagging on historical maps. *Proceedings of the 5th Annual ACM Web Science Conference*. ACM.

Isoda, Y. & Tsukamoto, A. &, Kosaka Y. & Okumura T. & Sawai M. & Yano, K. & Nakata, S & Tanaka S. (2010). Reconstruction of Kyoto of the Edo era based on arts and historical documents: 3D urban model based on historical GIS data. International Journal of Humanities and Arts Computing, 3:21–38, 2010.

Kahan, J., Koivunen, M., Prud'Hommeaux, E., & Swick, R. R. (2002). Annotea: an open RDF infrastructure for shared Web annotations. *Computer Networks*, *39*(5), 589-608.

Koivunen, M. (2006). Semantic authoring by tagging with annotea social bookmarks and topics. *Proc. of SAAW2006, 1st Semantic Authoring and Annotation Workshop, Athens, Greece.*

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the seventeenth conference on Hypertext and hypermedia*. ACM.

Overell, S., Sigurbjörnsson, B., & Van Zwol, R. (2009). Classifying tags using open content resources. *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM.

Pearson, A. (2006). Digitizing and analyzing historical maps to provide new perspectives on the development of the agricultural landscape of England and Wales. e-Perimetron, 1 (3).

Rumsey, D., & Williams, M. (2002). Historical maps in GIS. *Past time, past place: GIS for history*, 1-18.

Schroeter, R., Hunter, J., & Newman, A. (2007). Annotating relationships between multiple mixed-media digital objects by extending annotea. *The Semantic Web: Research and Applications*, 533-548.

Sigurbjörnsson, B., & Van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. *Proceedings of the 17th international conference on World Wide Web*. ACM.

Unsworth, J. (2000). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this. *Humanities Computing, Formal Methods, Experimental Practice" Symposium, Kings College, London*.

Witschas, S. (2003). Landscape dynamics and historic map series of Saxony in recent spatial research projects. In 21st International Cartographic Conference, pages 59–60, Durban, South Africa, 2003.

This research was funded by the DM2E project as part of the Open Humanities Awards.

Joined-Up Early Modern Diplomacy: Linked Data from the Correspondence of Thomas Bodley

Centre for Editing Lives and Letters
Dr Robyn Adams, and Jaap Geraerts

Project Report

Project Aim

This project proposed to analyze the dataset generated by ***The Diplomatic Correspondence of Thomas Bodley, 1585-97*** by producing visualizations of people and geographical locations mentioned in the letters for a new project, ***Joined-Up Early Modern Diplomacy: Linked Data from the Correspondence of Thomas Bodley***.

Summary

The Diplomatic Correspondence of Thomas Bodley 1585-97 is an online edition of the letters in English written between Bodley and his diplomatic network, completed at CELL in 2011. The bulk of these [930] letters were written during his long embassy to the Low Countries, where for nearly nine years he represented Elizabeth I in the role of English agent on the Dutch Council of State during the conflict between the United Provinces and Spain (1588 – 1597). Bodley was positioned at the centre of a correspondence network which included his political masters back home in England, the men responsible for the military activities of Elizabeth's troops, and other English personnel affected by the conflict, such as the Merchant Adventurers based in Middelburg.

The contents of the letters feature a wide range of information types, spanning military movements, political events, dynastic marriage negotiations, individuals' petitions, secret intercepted intelligence and ongoing patronage strategies between Bodley and his superiors. The letters reveal the multiple roles Bodley was required to perform, from standing firm in difficult negotiations on behalf of the queen with the Council of State, to forwarding petitions from supplicants based on the continent to prominent figures in England. As such, his correspondents represent a wide range of social hierarchy, from European royalty (Elizabeth I), through the English nobles heading up different aspects of the Low Countries campaign (Lord Treasurer Sir William Cecil, Lord Admiral Charles Howard), to individuals seeking information, restitution, repatriation or assistance (Captain Oliver Lambert) or making petitory requests (Richard Saltenstall).

New Users, Mining our own Data

During the transcription and encoding period of the ***Diplomatic Correspondence,*** the research team made the decision to enrich the metadata of the project by encoding each mention of the people and geographical locations featured in the letters.[1] While the network of correspondents is relatively small, the nature of the correspondence – describing events unfolding in Western Europe over nearly a decade - means that the tagged references to people and geographical locations are numerous and form a generous dataset. We have behaved as new users, and have mined this dataset to produce the visualizations for ***Joined-Up Early Modern Diplomacy***.

Our dataset derived from the Bodley correspondence is ideal for combining historical research and data visualization. As Ruth and Sebastian Ahnert note, 'Letters offer themselves to network visualization and analysis in a much more straightforward way than other forms of literature'.[2] An epistolary network rendered in visual terms depicts both the relationship between correspondents and the physical journey of the letters themselves, creating a 'material link' or trace between the nodes. A central feature of the project was to

1 As Lorna Hughes states, 'the key tool in the resource discovery of a digital dataset, or indeed any dataset, is the concept of metadata, or data describing data. Metadata is used to describe a record or an archived resource in such a way, and using such descriptors, that a researcher can easily discover that it contains information relevant to their researches.' 'Resource Discovery and Curation of Complex and Interactive Digital Datasets' in Chris Bailey and Hazel Gardiner eds. ***Revisualizating Visual Culture*** (Farnham: Ashgate, 2010), p.48.

2 Ruth and Sebastian Ahnert, 'Protestant Letter Networks in the Reign of Mary I: a Quantitative Approach', ***ELH*** (forthcoming).

interrogate the existing data in such a way that they might produce glimpses of patterns and behaviours by Bodley and his correspondents which were previously difficult to detect when tabulated textually.[3]

Report

This project falls within an interesting moment, digitally speaking. In the months leading up to our proposal to the Open Humanities Awards (March 2013), scholarly networks were buzzing with ideas and new examples of network visualization, infographics and data visualization. The CELL project team had been interested in novel methods of digital representation which stretched traditional scholarly boundaries, and we were keen to try something new with the data we already possessed (which was in the public domain available on Github). The metadata embedded in the ***Diplomatic Correspondence*** appeared to be ideal material with which to test how data visualization fits into the academic skillset alongside other high-end technical skills such as palaeographical expertise. Data visualization has proved a popular tool for scholars looking for an alternative method of assessing and analyzing data, networks and groups, and network theory has gained a firm foothold in digital humanities projects. We were keen to test various features of the technique, to explore a) what new information it could tell us about a dataset we were already familiar with, and b) the positive benefits or caveats of such an approach. To emphasize, we weren't interested only in generating visualizations from the data: we wanted to assess how useful were both the process and the results of those visualizations. As the ***Tooling Up*** team at Stanford have noted, 'We may tend to think of visualization as a finished product, as part of presentation, but it may be more useful to think of visualization as part of the research process'.[4]

After hiring a research assistant, Jaap Geraerts, who had both technical experience and knowledge of the historical context, we set about preparing the data for visualization. The first part of the process consisted of rationalizing the data between the two databases (a Microsoft Access and a MySQL database) in which the project data is stored. The MySQL database contains the data which can be viewed on the project website (the 'online edition' as it is defined), whereas the Access database includes a larger body of material, namely deriving from a preliminary census of all letters in which Bodley was mentioned, and letters written in foreign languages dating from the period in which Bodley served as an ambassador in The Hague. Because both databases have been used for the visualizations, it was important to make sure that the data stored in both databases was identical. The differences between the data in the databases were mostly detected via queries, after which the databases were manually updated. These modifications were also applied to the relevant XML-files, thus ensuring that the MySQL-database and the website were completely in sync with one another.

In order to optimize our dataset for the creation of visualizations, where there was an unknown author or recipient, we inferred these values, thus augmenting the existing data and expanding the possibilities for analysis. At a later stage in the project specific information was added to the existing data as well, (such as the country in which the places mentioned were located), as a result of which the correspondence between specific people could be analysed in terms of their national and geographical location. In a similar fashion the whole dataset could be examined (e.g. which countries figured most prominently in the correspondence).

Historical Data and Contemporary Software

When the process of rationalizing the data was nearing its end, we started to familiarize ourselves with the software we had selected to create the visualizations: GEPHI.[5] GEPHI is open source software which is mainly used for network analysis and the creation of network visualizations. The main advantages of using GEPHI, besides its being open source, is that the software is regularly updated - often based on the demands of its users - and that it is easy to use. Before we could start working with the Bodley dataset, however, we had

3 See Dan Dixon, who comments on methods of seeking patterns which 'would not be readily apparent to a human reader and require the brute force, or transformation, that computational methods bring which are usually difficult, boring, or physically improbable for human researchers to carry out', in 'Analysis Tool or Research Methodology: Is there an Epistemology for Patterns?' in David M Berry, ed. ***Understanding Digital Humanities*** (Basingstoke: Palgrave Macmillan, 2012), p.191.

4 'Tooling Up for Digital Humanities', Stanford University, <http://toolingup.stanford.edu/?page_id=1255>, accessed 21/03/14.

5 We worked with version 0.8.2.

to consider and assess our data - derived from historical sources - in relation to the capacity of the software platform to deal with it. This was probably one of the most interesting and informative parts of the whole project. For GEPHI is a standardized piece of software (as are most of the tools normally used in Digital Humanities projects), which on the one hand has the advantage that it can be used to work with various sorts of data fairly easily - people have set out to import various sorts of data ranging from social connections generated by Facebook to the epistolary network centered around the Roman lawyer Pliny the Younger.[6] However, the fact that GEPHI is generic software contains inherent drawbacks as well; one being that the historical data is often 'messier' than a neatly programmed piece of twenty-first-century software, as a result of which the complexity of the historical sources cannot always be fully captured by the software. Because of this, questions often need to be posed to the software in different ways than the traditional methods in which scholars have hitherto been trained.

The following examples will illustrate this point. Some of the letters which comprise part of the dataset were authored by more than one person or were sent to more than one recipient (there were also letters with several authors and recipients). This is a trait common to early modern epistolary networks. This was difficult to capture in GEPHI, for each node represents one author or recipient, while every edge, the line connecting the nodes, represents one letter. Because one edge can only connect two nodes, (at least in the current version of GEPHI), the fact that some letters had more than one author or recipient could not be visualized at all, and we had to proceed as if every letter had one author and one recipient (i.e. a letter with one author and two recipients in our dataset is seen as two letters in the data imported into GEPHI - one letter from the author to each recipient). Working with this off-the-shelf software thus required an additional amount of editorial intervention as decisions about the way we treated and proceeded with our data had to be made before we could proceed with creating the visualizations.

Another characteristic of the historical sources which has proved complicated to capture is the fact that letters in reality often were 'packets'; other documents were often enclosed with letters, e.g. copies of letters or more exotic objects such as maps. Sometimes recipients were asked to distribute portions of the material that was sent along with the letters they received to other people, and thus new links were created that existed external to the epistolary network (in the sense that these relationships were not directly forged by one person sending a letter to another person, but were transmitting information derived from within the letters included in the epistolary network). We have called these people ***transmission agents***. It would have been extremely interesting to include these multifaceted and nuanced relationships in the network. However, the different features of the various links and activities of letter-writers and transmission agents are difficult to visualize using GEPHI, for although attributes can be added to the edges connecting the nodes, only one edge can connect two nodes; hence only one sort of relationship can link two nodes. Because of these limited options to treat nodes and edges which have different attributes, we found it difficult to incorporate a range of relationships into one network and one visualization.

We found a cognate problem in trying to analyze a 'multimodal' network; the in-house GEPHI algorithms do not distinguish between various types of edges - although one can work around this by filtering a network (reducing the network to one type of node/edge, i.e. one type of relationship), then running the statistical analysis over this filtered network, and comparing the results. One could circumvent these limitations by constructing separate networks, each of them based on a specific sort of relationship, and to compare the analysis of these separate networks. The limitations of the available software made us realize that the complexity of a dataset often cannot be captured in one visualization, for the specific characteristics of the visualization (or of the software used to create it) might not allow for the flexibility necessary to incorporate all aspects of the (historical) sources. In addition, there is often simply too much information, and including this would create an incredibly dense image, at once reducing the added value of visualizations (i.e. quickly discerning significant patterns). We therefore set out to think about which visualizations are best suited for depicting certain aspects of the dataset and the way in which these visualizations complement each other.

6 E.g., Sarah Joy, 'Using Netvizz & Gephi to Analyze a Facebook Nework', <https://persuasionradio.wordpress.com/2010/05/06/using-netvizz-gephi-to-analyze-a-facebook-network/> and 'Visualizing Historical Networks: Pliny Letters', Harvard University, <http://www.fas.harvard.edu/~histecon/visualizing/pliny/index.html>, both accessed 21-3-2014.

This 'tension' between our historical data and our selected software made it necessary to modify the CSV-files that are used to import data into GEPHI. The files are basically lists of nodes and edges, the latter list consisting of a 'source' node and a 'target' node, and because of the discrepancies between our data and the data-format used by GEPHI, we needed to manually update these lists (in the case of multiple authorship or recipients) to ascertain that GEPHI processed the data in the correct way. So, besides the methodological and editorial decisions which had to be made, importing the data into this software required additional manual work. We stress this point because it is vital to account for the labour required prior to the actual creation of the visualizations. For although the IT-tools commonly used in Digital Humanities research, (i.e. data-mining, (network) analysis and the creation and output of visualizations), can add considerable value and extend scholarly research into other domains, to achieve this scholars are required not only to gather the corpus of material and data (activity which demands a host of expertise in itself) but often need to manipulate, disambiguate or modify the data before it can be processed by computer software.[7] It is essential that all these processes be executed in a methodologically robust way. Thus, research undertaken in Digital Humanities relies on the successful marriage of traditional research methods with a sound understanding and application of IT-technologies.

Visualizations

All visualizations created in the project will be incorporated into the ***Diplomatic Correspondence*** website.

The aim of this project was speculative in the first instance, i.e. the visualizations were not intended to support existing research questions but rather to detect new patterns and frameworks for analysis. Our interest lay in the whole process of creating visualizations: of seeking the connective tissue which comprised Bodley's correspondence and to enhance our understanding of the advantages and pitfalls of working with IT-tools and visualizations. We wanted to analyze the dataset as a whole as well as the relationships between correspondents. As such, the primary impetus behind this project was not the visualizations themselves, but rather the lessons learned from thinking about visualizations and the process of relating our data to the software used to produce them. As well as the realization that additional editorial intervention was necessary at different times in the process, we gained new insights into the possibilities and limitations of data visualization by testing the boundaries of the software currently available.

The visualizations were created by using a combination of various programs such as GEPHI, Adobe Illustrator, Inkscape, and Microsoft Visio. The data on which the visualizations are based was generated by executing queries in the Access and the MySQL databases. After the results from the queries were put into the right format so that they could be imported into GEPHI, the work on the actual visualizations could finally commence.

Constraints

One of the limits of the majority of recent network visualizations is that they can contain only one layer of information - i.e. only one sort of relationship is depicted. Users have tested the limits of the various available visualization software platforms in order to assign various types of nodes,[8] but so far the development of 'multimodal' networks as they are sometimes termed is still in its infancy. The limitations of networks which include only one type of relationship are obvious, for although we can depict a correspondence network, other relationships which linked correspondents to each other cannot be visualized, that is, it is difficult to capture the nuances of all these relationships in one visualization.[9] This project has exposed the boundaries of the visualizations currently enabled by GEPHI. However, by introducing another layer of information, namely the

7 This is not always the case: sometimes the data is available to be mined and analysed, but the specific research questions require the software to be updated or altered. See, e.g. Joris van Eijnatten et al., 'Big data for global history: the transformative promise of Digital Humanities', ***Low Countries Historical Review*** 128:4 (2013) 55-77.

8 See, e.g. "Visualizing Historical Networks: People and Institutions', Harvard University, <http://www.fas.harvard.edu/~histecon/visualizing/graphing/people.html>, accessed 21-3-2014.

9 E.g., Ruth and Sebastian Ahnert, 'Protestant Letter Networks in the Reign of Mary I: a Quantitative Approach', figure 1.

metadata of the places and people mentioned in the letters, we have managed to achieve positive results. (This was easier said than done, for the introduction of a new layer of information required the MySQL database to be modified, as well as the CSV-files used to import information into GEPHI).

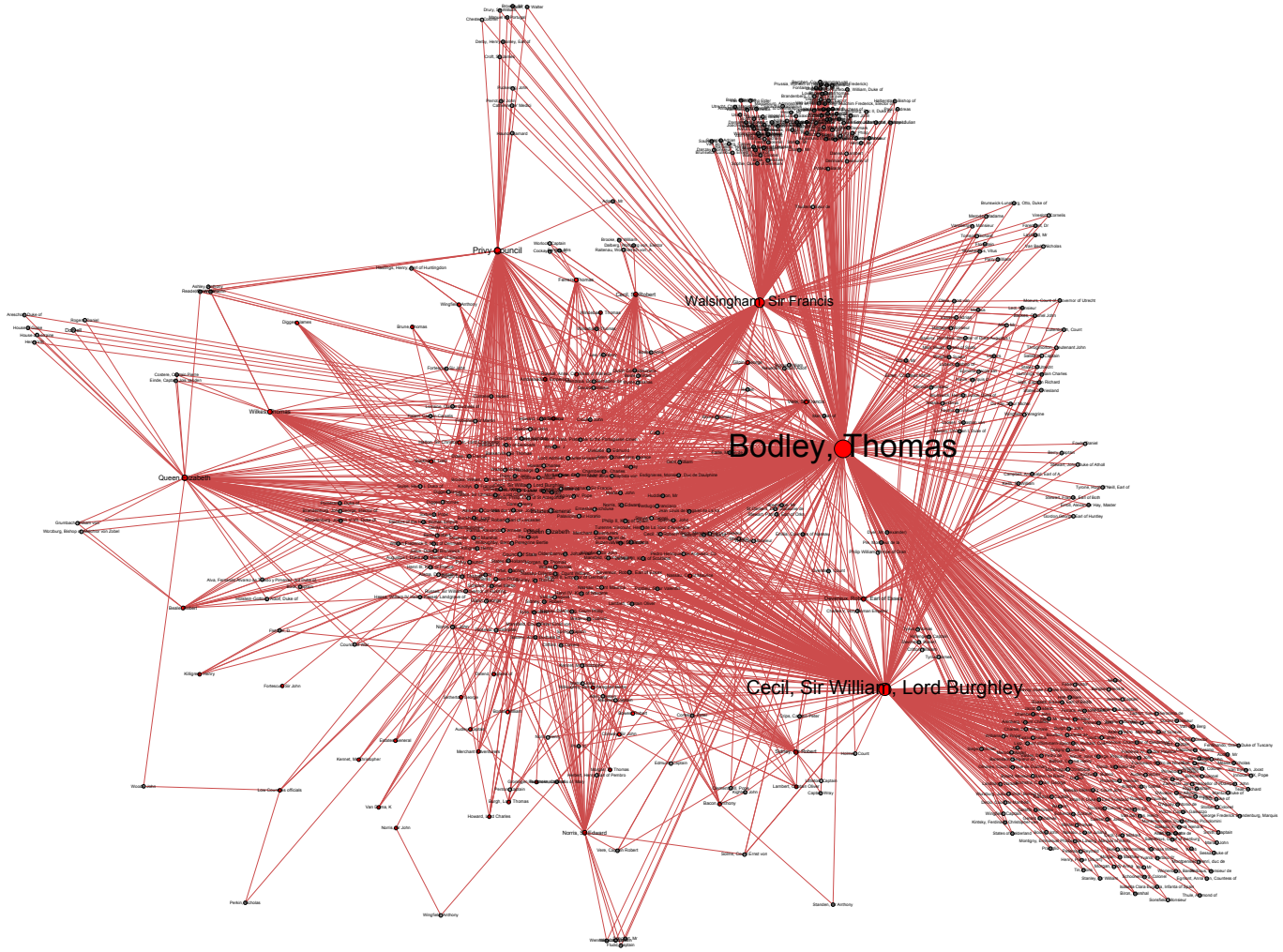


Figure 1. People: entire network 1 (hairball)

The Visualizations in Detail

Importing the comprehensive metadata to GEPHI results in an extremely dense image which has required manual modification in order to produce a visualization in which the relationships are clear. Because of the density of this image, as an overview it is not extremely useful. We could not resort to algorithms which can normally be used to improve the 'readability' of networks, (such as the Force Atlas algorithm),[10] because these are fine-tuned to work with straightforward one-to-one relationships. Introducing another layer of information (the people and geographical locations) creates a one-to-one-to-one relationship (between the author, the information mentioned, and the recipient, respectively). However, although at first instance the scholarly advantages of such a visualization are difficult to grasp, when zooming in and looking at the specific correspondence between two people, useful patterns start to emerge. It became immediately apparent, for instance, that there were a number of Scottish noblemen that were only mentioned in the correspondence between Bodley and Sir Robert Devereux, earl of Essex.

10 'New Tutorial: Layouts in Gephi', <http://gephi.org/tag/force-atlas/>, accessed 21-3-2014.

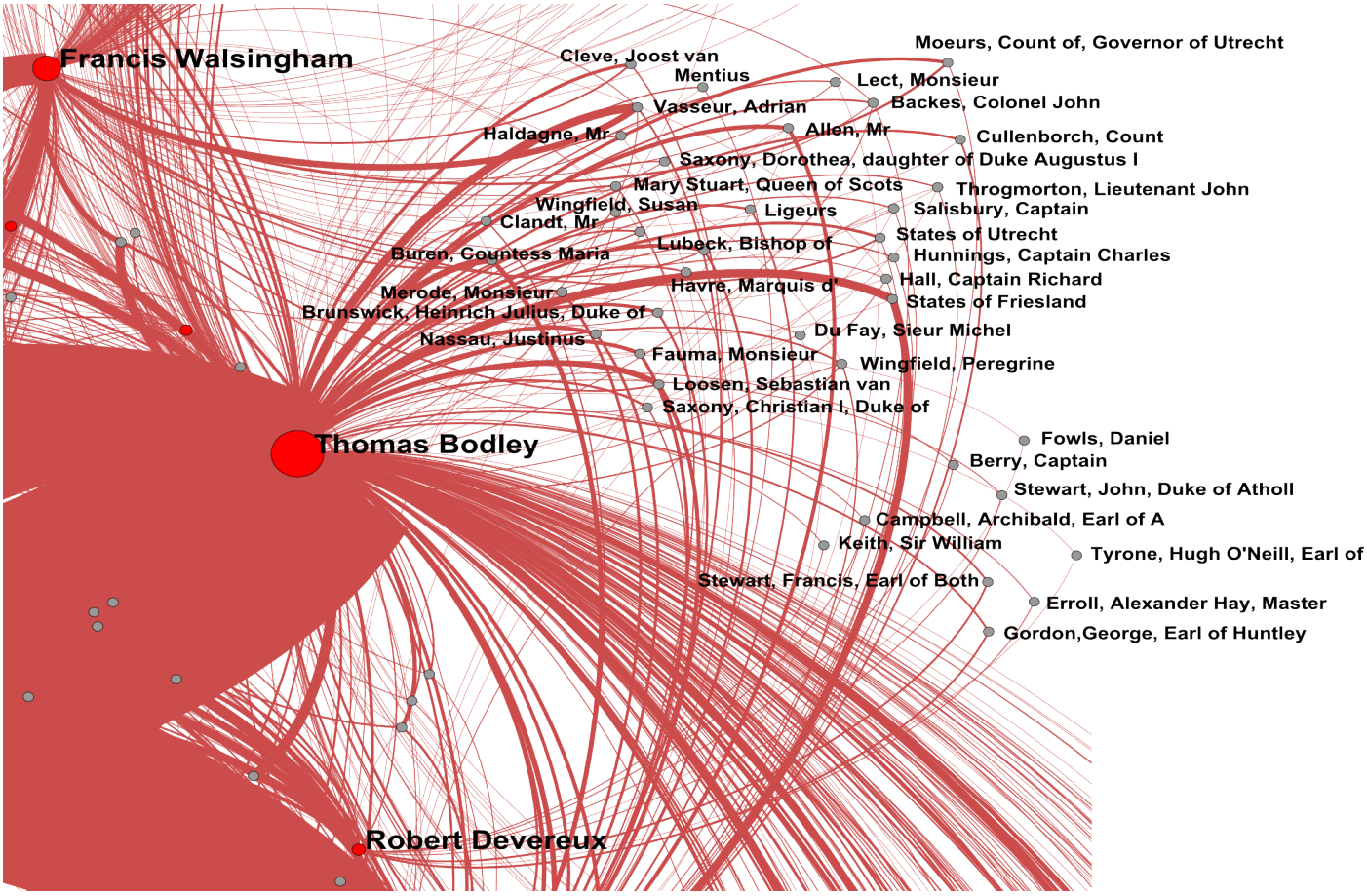


Figure 2. People: Bodley and Devereux (zoom)

To increase the readability of this visualization we applied filters to the network, as has been done in the previous image, where the in-and-out degree (the number of directed edges which go to or from a node, or the number of connections associated with a node) is set to 11. This setting of 11 filters out a number of nodes (people mentioned as well as authors/recipients) which did not meet this criterion leaving a much less dense network, and providing a concise overview of the most important correspondents and the people they mentioned in their letters.

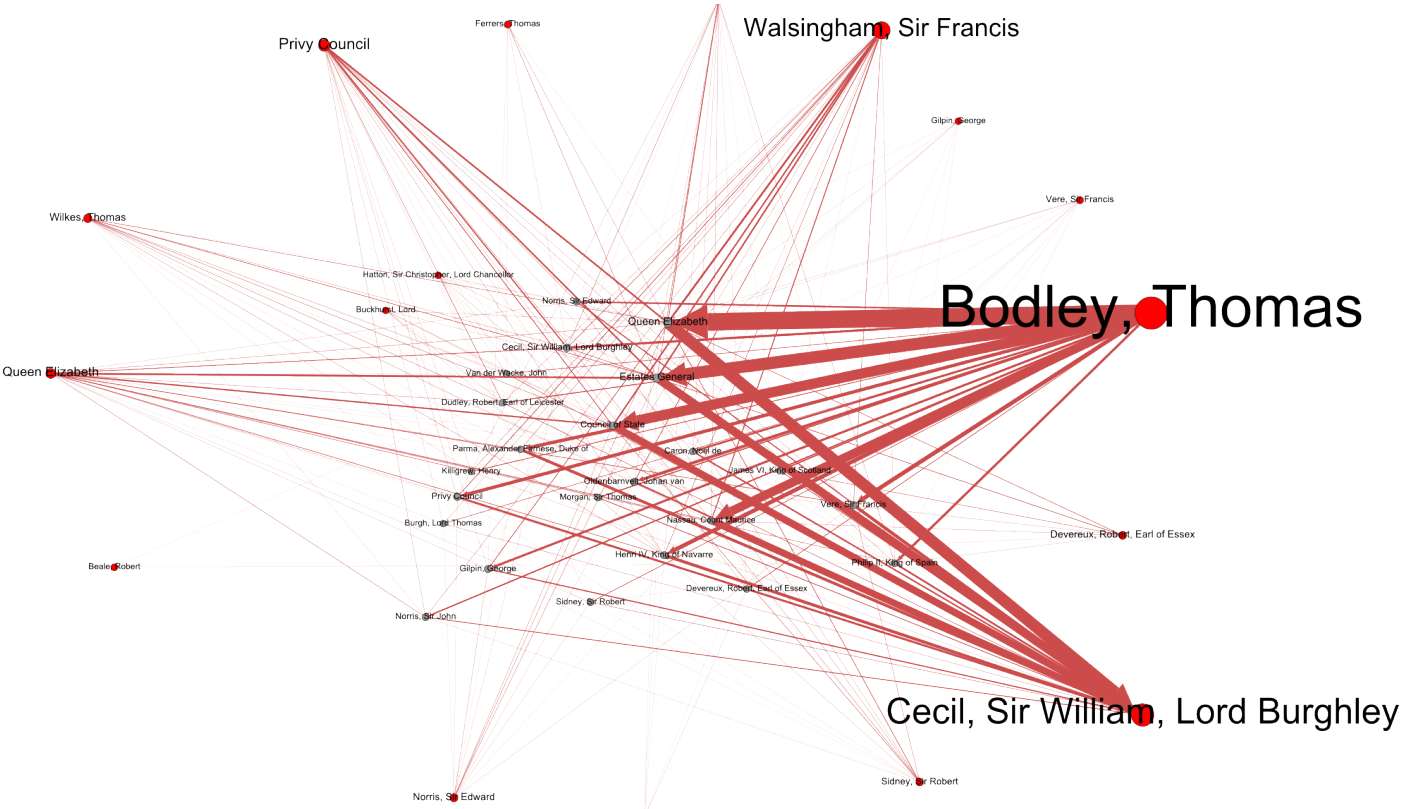


Figure 3. People: (filter with 11 degree)

As patterns within the dataset are detected by using different kinds of (statistical) analysis, we soon realized that, due to the limits of specific types of visualizations, it was better to depict the outcomes of these analyses using different ***kinds*** of visualizations. The dataset as a whole can be analyzed, for instance, by looking at the frequency with which places and people were mentioned, showing the degree to which the correspondence covered a wide or rather small range of people and geographical locations. This can be done quite simply by employing bar graphs.

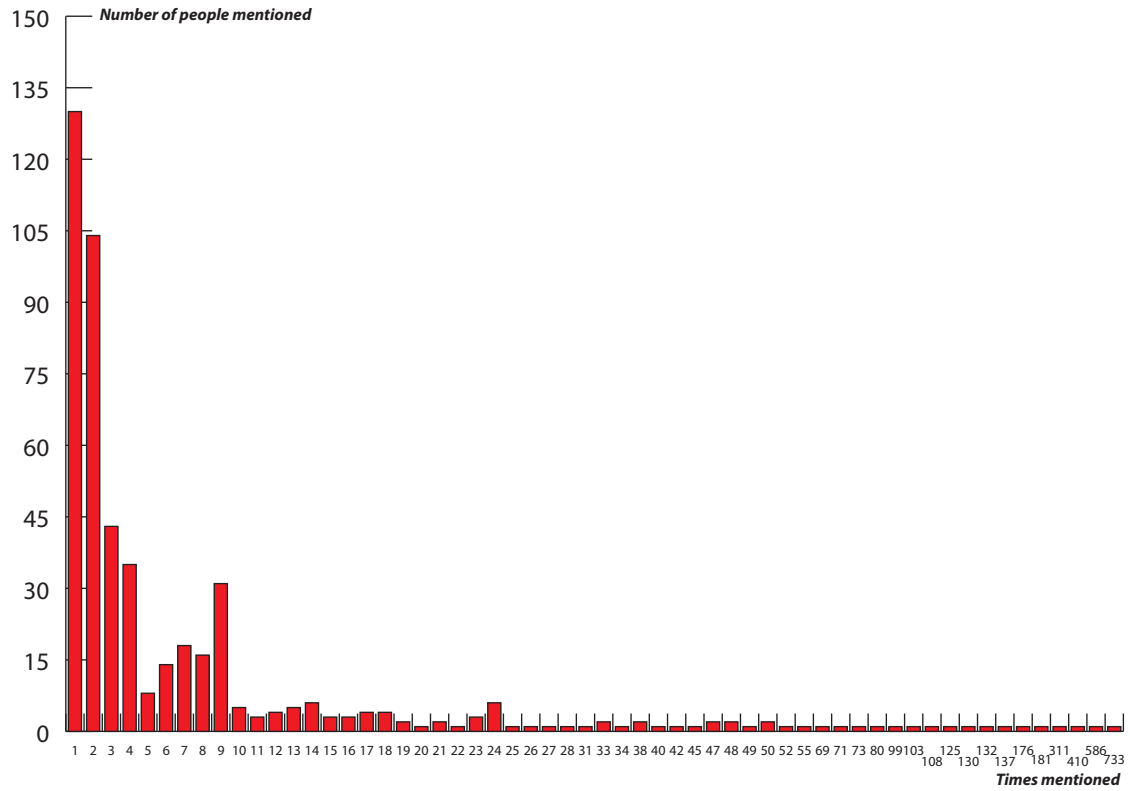


Figure 4. People: frequency of mentions (bar chart)

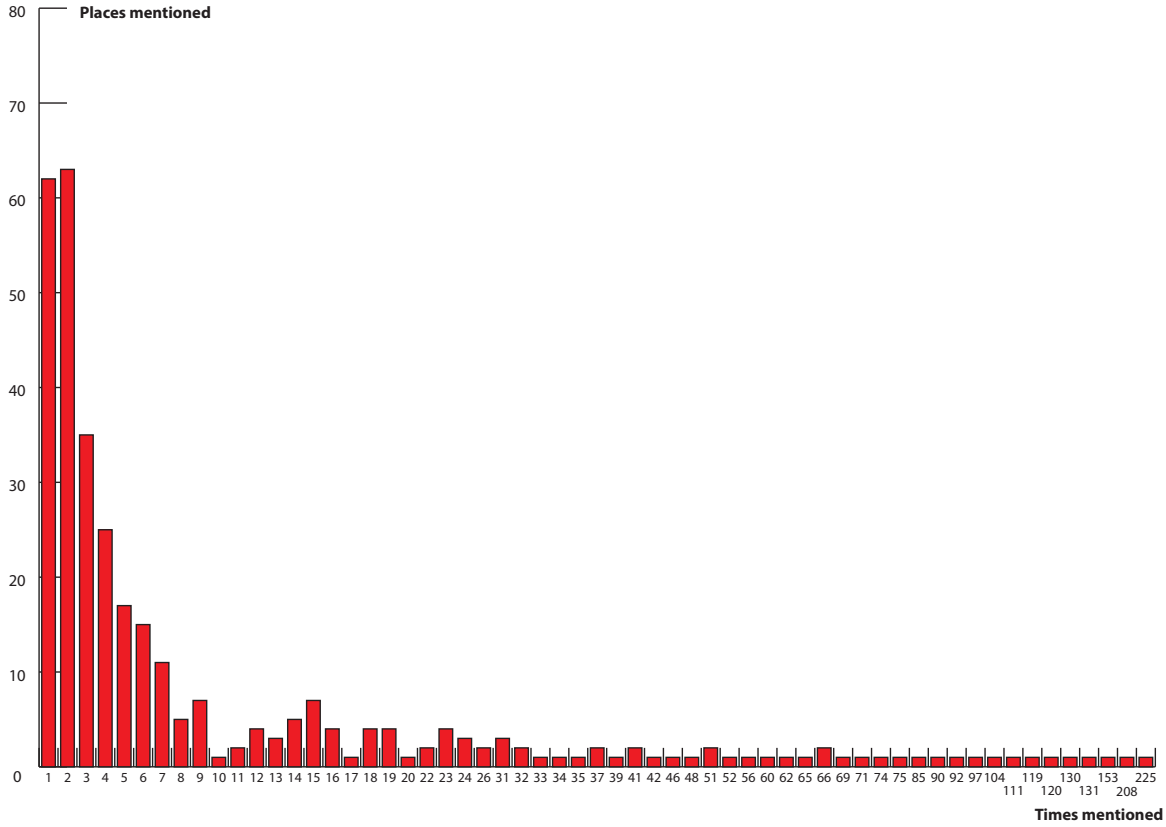


Figure 5. Places: frequency of mentions (bar chart)

Both charts show that most of the places and people were mentioned just one or two times, and that there only were a very limited number of places and people which were mentioned more than a hundred times. The correspondence, although covering a wide range of people and places, clearly centered on a select number of them (understandable, considering Bodley's mission in the Low Countries, focused on western European affairs), and such visualizations, which give a general impression of the correspondence as a whole, give the scholar a lead on which patterns might benefit from further, in-depth analysis. More detailed information can be gathered by zooming in at the correspondence between two people, as has been done in the following bar graph.

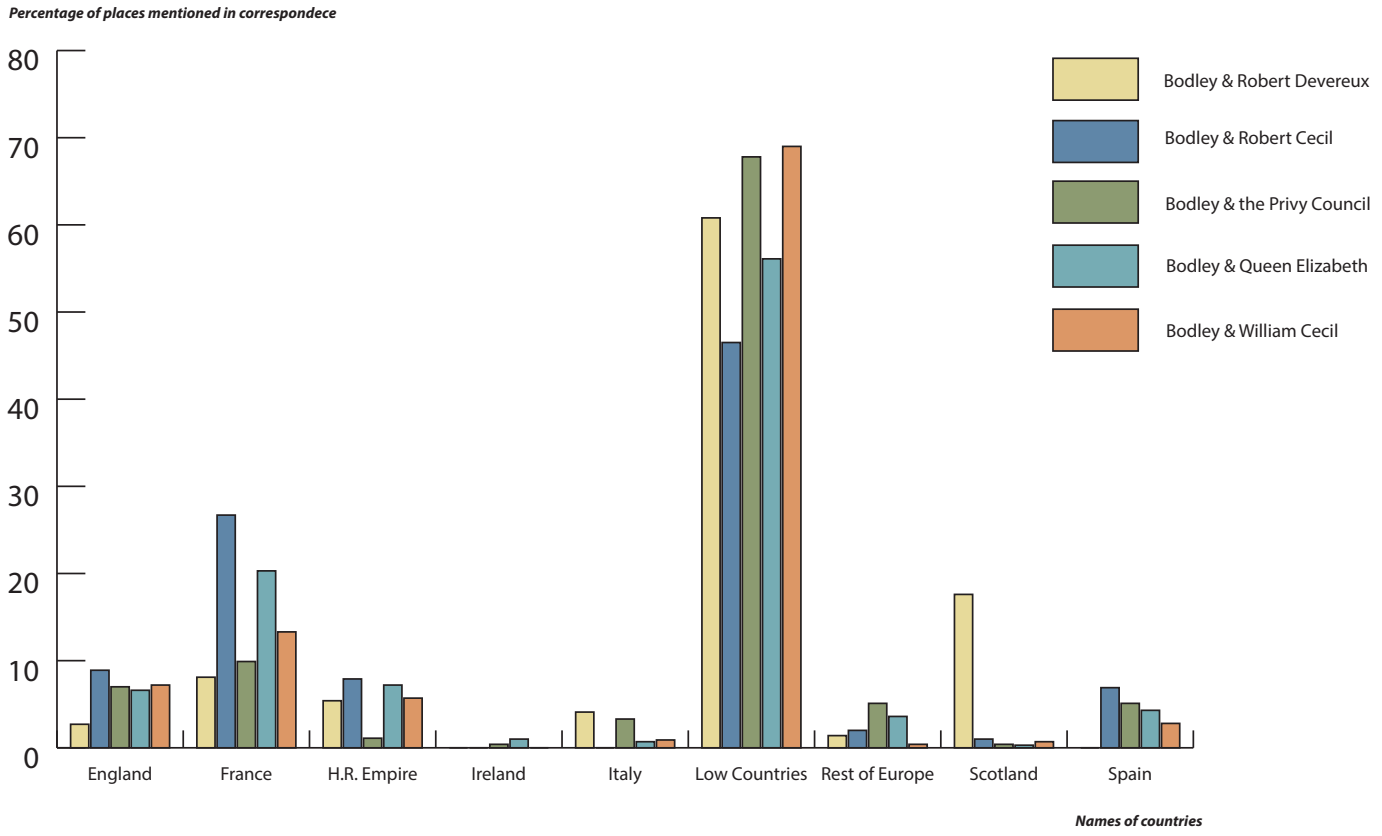


Figure 6. Places: mentions in individual conversations

By adding information to the place mentioned (the country in which these places were located), this visualization quickly shows which countries the various conversations centered. Compared to his correspondence with other people, Bodley and Sir Robert Devereux devoted a relatively large part of their letters discussing places in Scotland (which was not a surprise considering the patterns visible in other visualizations), while Bodley and Sir Robert Cecil often spoke about places in France. These are just a couple of examples of how augmenting the existing data with additional information and using a combination of different visualizations, each based on its own subset of data, can work together to create new patterns and relationship networks, that are hard to detect without the assistance of IT-tools.

A final example of some of the visualizations that have been created are the two SDL-diagrams below. The acronym SDL stands for Specification and Description Language and these diagrams are normally used to visualize aspects (e.g. actions) of a particular process taking place within a system (e.g. a computer program). We have, however, used these diagrams for creating visualizations which incorporate a large subset of the information that is to be found in the dataset, including actions the author of the letter required from the recipient, and the relationships that were established as a result of those requirements. The diagrams are based on two case studies, namely the sieges of Geertruidenberg (1589-93) and Groningen (1594) and their immediate aftermath, which generated a select sequence of letters. The diagrams make clear that writing and receiving letters was only a part of the process, as often recipients were asked to pass on to third parties (part of) the information included in the letters or documents enclosed with the letters. The visualizations show that letters did not only create a relationship between the author and the recipient, but rather forged a number of links, thus expanding the network beyond the standard binary epistolary structure, while also providing understanding of the way the information flowed through this network. For useful as they are, the network visualizations which depict the connections between the authors and the recipients only show a part of the organic and brittle process of gathering and disseminating early modern information, and omit the people who were closely related to these networks as transmission agents.

The diagrams, developed to highlight the complexity of the network (and the way information was disseminated through the network and beyond), require some explanation, and the meaning of the various icons are given in this overview.

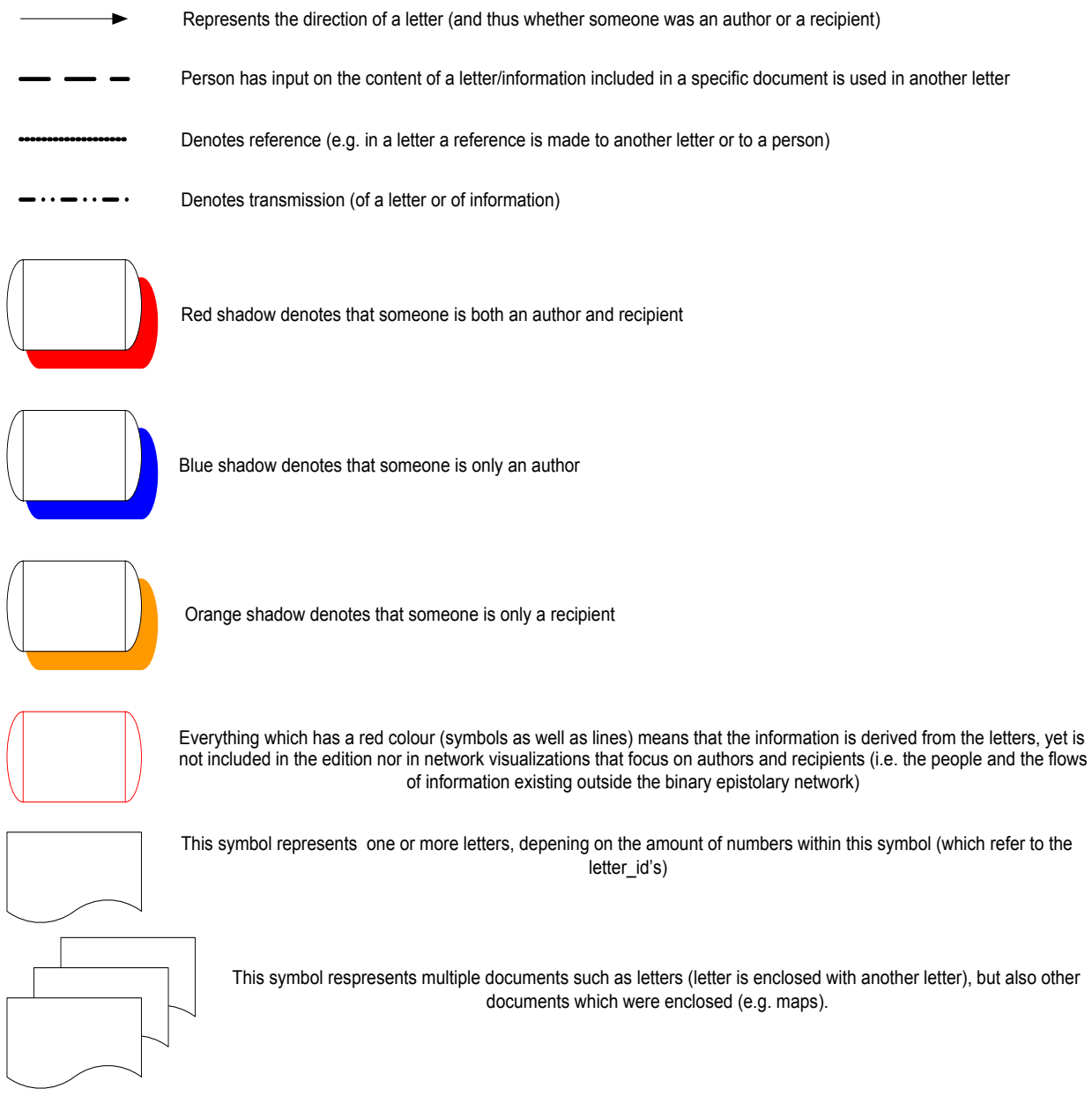


Figure 7. Key to SDL diagram

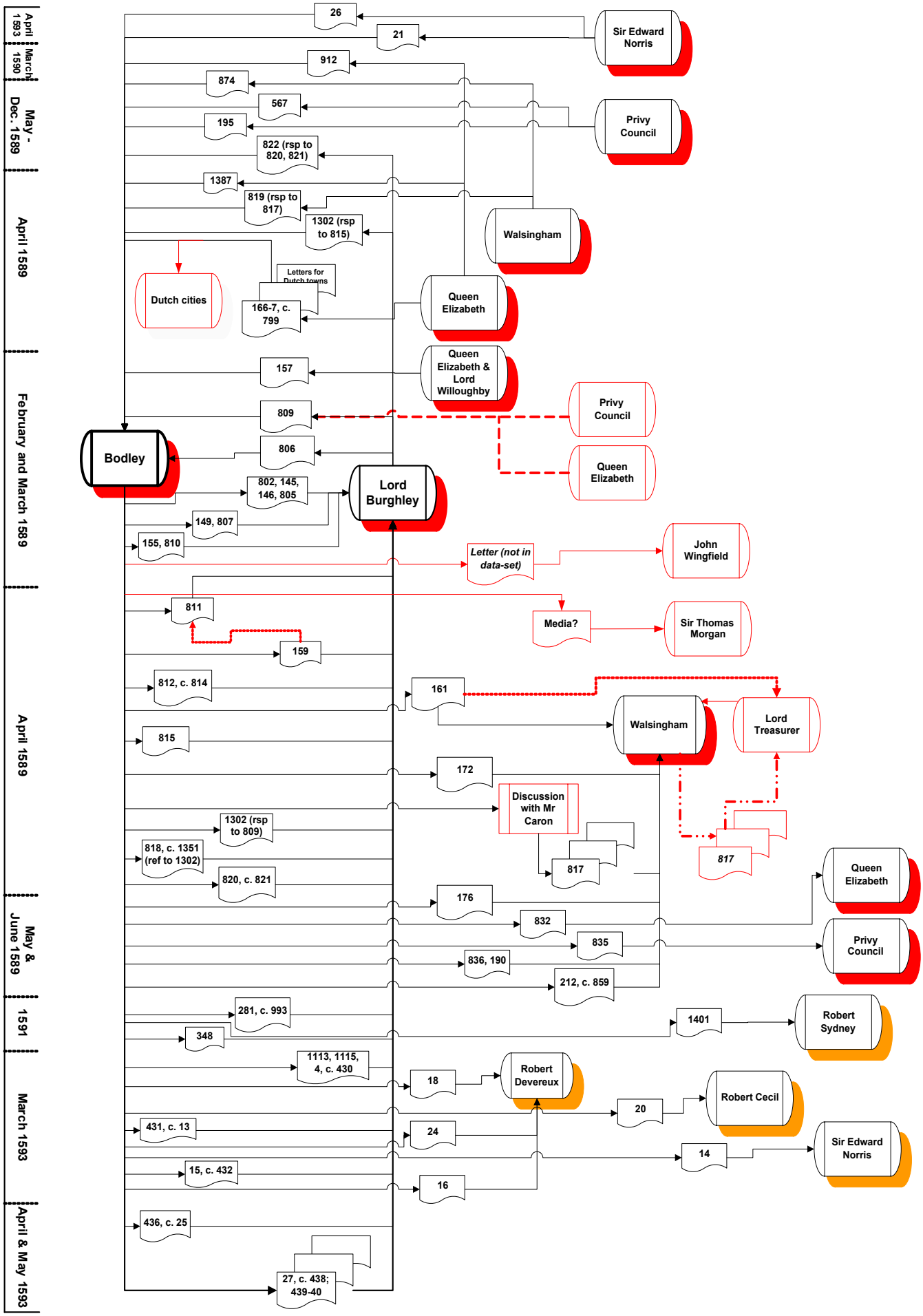


Figure 8. Case study 1: Geertruidenberg (SDL)

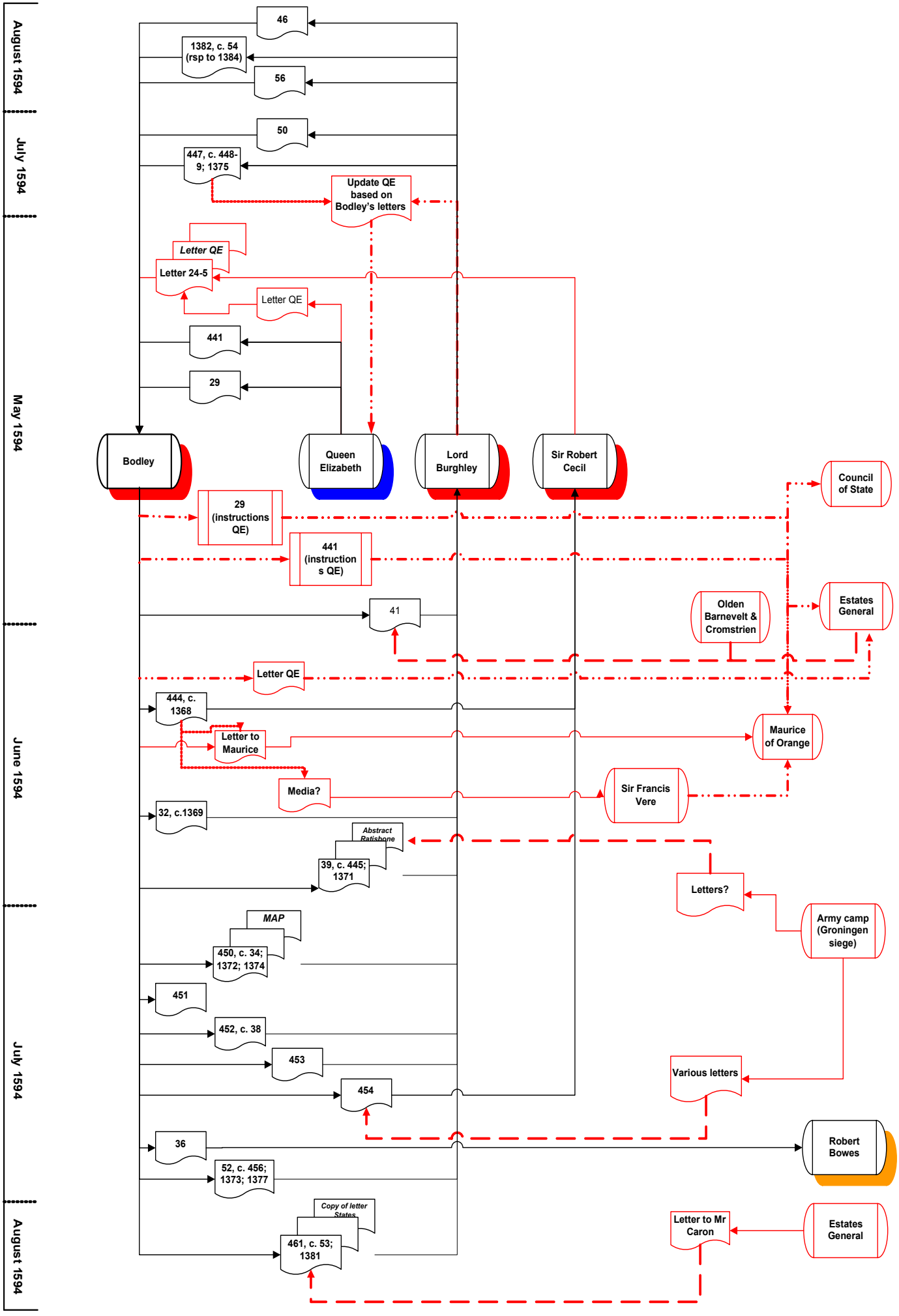


Figure 9. Case study 2: Groningen (SDL)

Let us survey a couple of examples included in the diagram which depicts the correspondence regarding the siege of Groningen. In his letter to Sir Robert Cecil (June, 6, 1594; letter id. 444), Bodley mentions that he received Cecil's letter of May 24, 1594, which is no longer extant and not a part of the dataset, (and therefore not included in the binary epistolary structure). Enclosed with Cecil's letter was a letter written by Queen Elizabeth, which Bodley duly brought to the States General 'the next day after [he had received Cecil's letter], and they promised to answear, when according to their custome, they shall have taken some time to deliberat upon it'.[11] Bodley thus not only performed some actions, leading to links which are normally not included in the binary networks consisting of authors and recipients, but he also provided information about his actions and, in this case, the expected results. In the same letter Bodley tells that he tried to convince to Dutch authorities to 'arme as many shippes', and he mentions that he had written a letter to stadholder Maurice of Orange and also 'requested Sir Francis Vere, to urge him [=Maurice] to it very earnestly' (it is unknown whether Bodley did so by via a letter, hence the speculative note 'Media?'). On 14 July, 1594, Bodley wrote that he had 'lettres from the Campe at this very instant', and he included some information he derived from these letters in his letter to Burghley (letter id. 0454). By including such links in this diagram, we are able to show the larger scope of the network and, specifically, the ***routes*** on which this information flowed through the network.

Outcomes

This project to analyze the dataset and to generate visualizations has proved enormously helpful in enabling us to detect patterns of communication and relationships in Bodley's correspondence network. It is clear that the process of using data visualization as an aid to historical research can identify additional and alternative routes of enquiry. For instance, only through close textual analysis would we have perceived that Bodley's correspondence with Sir Robert Devereux focused heavily on Scottish affairs, whereas it is immediately perceptible through Figure 2. People: Bodley and Devereux (zoom). Tracking the routes of transmission of specific sequences of letters – those concerning Geertruidenberg and Groningen – using the SDL diagram reveals with clarity the fluid and organic nature of a cross-channel early modern correspondence network. These examples demonstrate how valuable the visualization process can be to assist the scholar in interrogating a corpus of material with large or detailed questions.

However, there are some equally useful caveats to consider when deploying or using network visualizations in historical research which we have recognized during this learning process. It is essential that appropriate care and attention be paid to contextualizing the resource created. We have encountered numerous examples of visualizations where insufficient attention has been devoted to providing a suitable background and historical context to the results depicted. The lack of sufficient information given often results in decreased understanding of what exactly the visualization is trying to convey. Of course, too much information may have the result of data duplication; but if the context provided is robust enough (by means of an introductory summary, or running commentary with each visualization, for example), then swift perception of the data and patterns will occur. The confluence of interest in data visualization and infographics means that there is often positive overlap between well-constructed visualizations of networks and aesthetic presentation. However, problems can occur when the visualization is beautifully presented and artistically designed but the reader has little contextual information with which to understand the image.

It is also fundamentally important to take into account the individual context for each dataset. It is not enough to stumble across an open source dataset and begin generating visualizations; care must be paid to understanding the conditions surrounding the production of the data, the personnel and the (archival) material involved. In our case, certainly, the editorial questions raised by the material (what constitutes an early modern letter? what are the contemporary circumstances which prompt the writing of letters? how can we visualize the specific role of the transmission agents?) demanded that we approach the production of the visualizations with a sensitive eye to the historical context of the correspondence.

11 'The Diplomatic Correspondence of Thomas Bodley', CELL, <http://www.livesandletters.ac.uk/cell/Bodley/transcript.php?fname=xml//1594//DCB_0444.xml>, accessed 21-3-2014.

Overall, this small project has provided us with some valuable insights. Of primary interest are the patterns and connections which were previously imperceptible without significant and time-consuming textual analysis of what is a substantial corpus of material. But it has been additionally positive to experience first-hand the value of having collected the metadata of people and places at the point of transcription. Without that extra subset of data these would have been a very bland set of network visualizations, and Bodley's small number of correspondents would have been perceptible without computational methods. The extra depth provided by this metadata has provided an alternative (and interesting!) route of research, and demonstrates that projects such as the ***Diplomatic Correspondence*** are value-added when the time is taken to take the data-collection stage to a higher level. Our next task will be to investigate in detail the fascinating patterns and routes of enquiry generated by the visualizations.

Linking Early Geospatial Documents, One Place at a Time: Geo-Tagging Texts and Maps with *Recogito*

Rainer Simon, AIT Austrian Institute of Technology, rainer.simon@ait.ac.at
Elton Barker, The Open University, elton.barker@open.ac.uk
Leif Isaksen, University of Southampton, l.isaksen@soton.ac.uk
Pau de Soto Cañamares, University of Southampton, p.desotocanamares@soton.ac.uk

Recogito *is a Web-based tool for the structured annotation of place references in texts and images. As part of the Open Humanities Awards 2014, we held two "hackathon"-like workshops, where a mixed audience of students and academics of different backgrounds used* Recogito *to annotate literary texts from the Classical Latin and European Medieval period, as well as Medieval Mappae Mundi and Late Medieval maritime charts. At the end of the day, participants had added several thousand contributions, all of which are now openly available for download and further re-use. The resulting data can be used, for example, to "map" and compare the narrative of the texts, and the contents of the maps with modern day tools like Web maps and GIS; or to contrast documents' geographic properties, toponymy and spatial relationships. Contributing to the wider ecosystem of the "Graph of Humanities Data" that is gathering pace in the Digital Humanities (linking data about people, places, events, canonical references, etc.), we argue that initiatives such as this have the potential to open up new avenues for computational and quantitative research in a variety of fields including History, Geography, Archaeology, Classics, Genealogy and Modern Languages.*

1. Background: the Pelagios Project and SEA CHANGE

Pelagios[1] is a community-driven initiative that facilitates better linkage between online resources documenting the past, based on the places that they refer to. Our member projects are connected by a shared vision of a world – most eloquently described in Tom Elliott's article *'Digital Geography and Classics'* [1] – in which the geography of the past is every bit as interconnected, interactive and interesting as the present. Each project represents a different perspective on our shared history, whether expressed through text, map or archaeological record. But as a group we believe passionately that the combination of all of our contributions is enormously more valuable than the sum of its parts.

The goal of Pelagios' current project phase ("Pelagios 3", funded by the Andrew W. Mellon Foundation) is to annotate, link and index place references in digitized *Early Geospatial Documents* – documents that use written or visual representation to describe geographic space prior to 1492. Through a series of six thematic work packages, Pelagios 3 will work with documents from the Latin, Greek, European medieval, maritime, as well as early Islamic and Chinese tradition. *Recogito* is a Web-based tool we developed specifically for use within the project team, to facilitate this work. However, the potentially unlimited number of documents to which our methodology would be suited means that establishing and honing community-based approaches will be essential in order to scale it beyond the pre-modern era.

The Open Humanities Awards have provided us with an impetus for trialing *Recogito* with a wider audience: under the title SEA CHANGE,[2] we held two public geo-annotation workshops with a mixed audience of students and academics of varying backgrounds (geography, history, engineering, and archaeology). Our primary goal was to explore the potential of *Recogito* as a tool for crowdsourcing and collaborative geo-annotation, but we were also interested in how and if a workshop format such as this is a suitable way to engage with a wider audience, and as a means to build community.

---

[1] http://pelagios-project.blogspot.co.uk
[2] **S**ocially **E**nhanced **A**nnotation for **C**artographic **H**istory **A**nd **N**arrative **Ge**ography, http://dm2e.eu/open-humanities-awards-round-2-winners-announced/

2. Recogito

Recogito features several work areas (see Fig. 1), each dedicated to different stages of the geo-annotation workflow: an image annotation area to mark up and transcribe place names on map or manuscript scans, a text annotation area to demarcate place names in digital text, and a geo-resolution area, where the identified (and transcribed) place names are mapped to a gazetteer (and, thus, to geographical coordinates). *Recogito* also provides basic features for managing documents and their metadata, as well as functionality for viewing and downloading annotation data and usage statistics. Editing functionality is limited to registered users. However, data downloads and basic overview information is also available for access to the public. Our own production instance of *Recogito* is hosted at http://pelagios.org/recogito. The tool as such, however, is Open Source software (available from the Pelagios project's GitHub repository http://github.com/pelagios/recogito), which makes it possible to set up additional instances of *Recogito* for personal or institutional use.

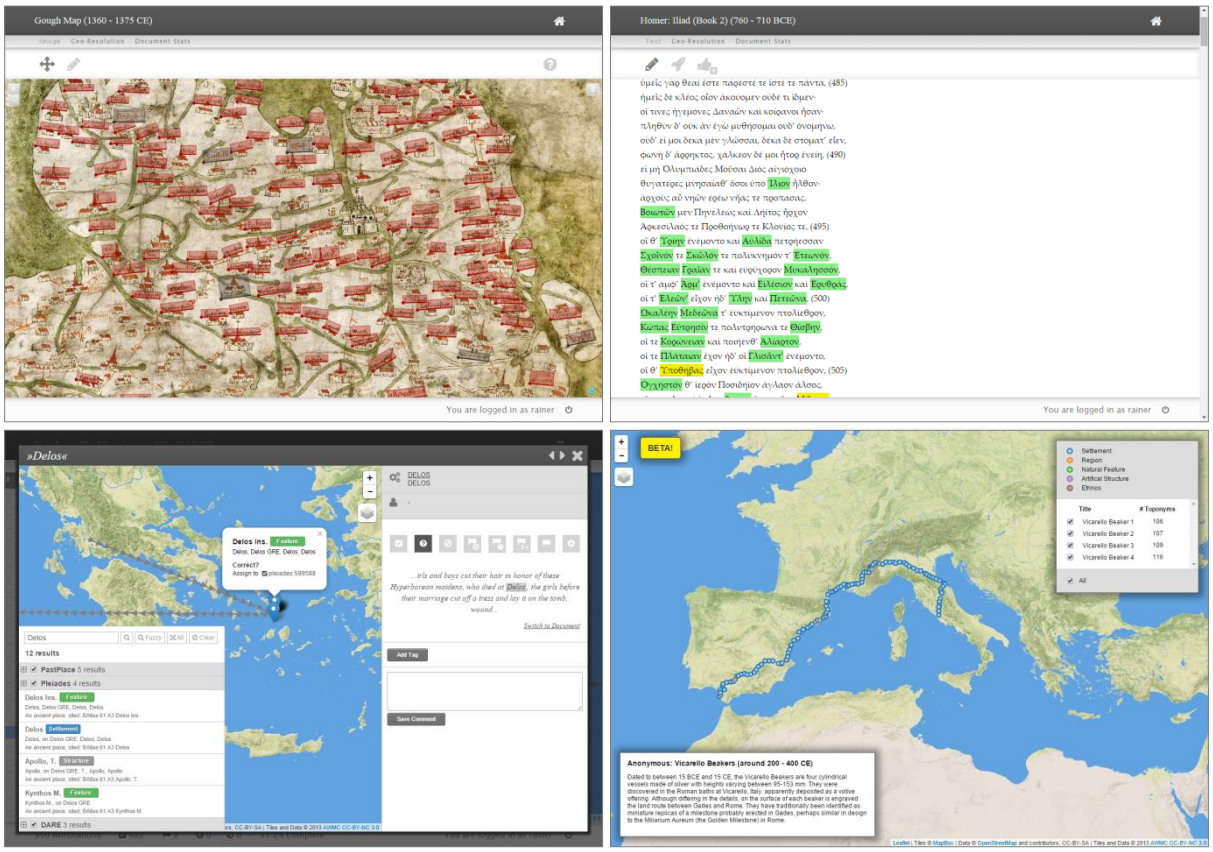


Fig.1 Recogito work areas: image annotation (top left), text annotation (top right), geo-resolution (bottom left), public map (bottom right).

3. Annotation Workshops

Our two workshops took place on October 31, 2014 at the Heidelberg University Institute of Geography, and on December 4, 2014 at the University of Applied Sciences Mainz. We started both days with a brief introduction to the goals and background of Pelagios, and a short tutorial of how to use Recogito's different work areas. (A written beginner's tutorial is also available online at http://pelagios.org/recogito/docs.) For each workshop, we defined a general thematic scope, and prepared material for annotating accordingly: Classical Latin texts and medieval maps for Heidelberg; Medieval travel writing and pilgrimage itineraries, and medieval nautical charts for Mainz. Beyond that, however, participants were free to choose which documents they wanted to work on, and which tasks they would focus on (tagging, transcribing, mapping toponyms to gazetteer records). Group sizes were roughly equal in both workshops, with 27 users in Heidelberg and 22 in Mainz.

After the introduction, we dedicated about 2 ½ hours to annotation work. The afternoon session, we used as a more open space for hands-on exploration. We wanted to get the audience thinking about the question: *"now*

that we have annotated our documents, what can we do that we couldn't do before?" As a concrete example, we prepared a tutorial which walked the audience through the steps necessary to download data from *Recogito* and analyze it further in QGIS (an open source Geographic Information System). This way, they could e.g. explore a medieval travel itinerary, and match the rate of stops and their different types against a 3D terrain model, pondering about the time taken – and the hardships endured – by travelers in the 4th century AD during their journeys. In the Mainz workshop, where part of the audience had an engineering background, we additionally prepared a short "hacking tutorial" consisting of small programming tasks that demonstrated how to re-use annotation data to create Web maps, timelines or network graphs, using JavaScript as a programming language.

3.1 Results Heidelberg

The quantity of contributions made by our participants greatly exceeded our expectations: on the first workshop day (Fig. 2), we recorded a total of 6,620 contributions, associated with 51 different documents (19 text documents, 8 of which were in Latin; and 32 map scans). Four participants even made it into our all-time top-10 list, which means that they managed to make more than 645 contributions in that morning session. The contributions consisted of approximately 2,650 place name identifications in text, 2,500 place name identifications on maps, 830 map transcriptions, 140 gazetteer resolutions and about 490 other actions, such as corrections, deletions or comments.

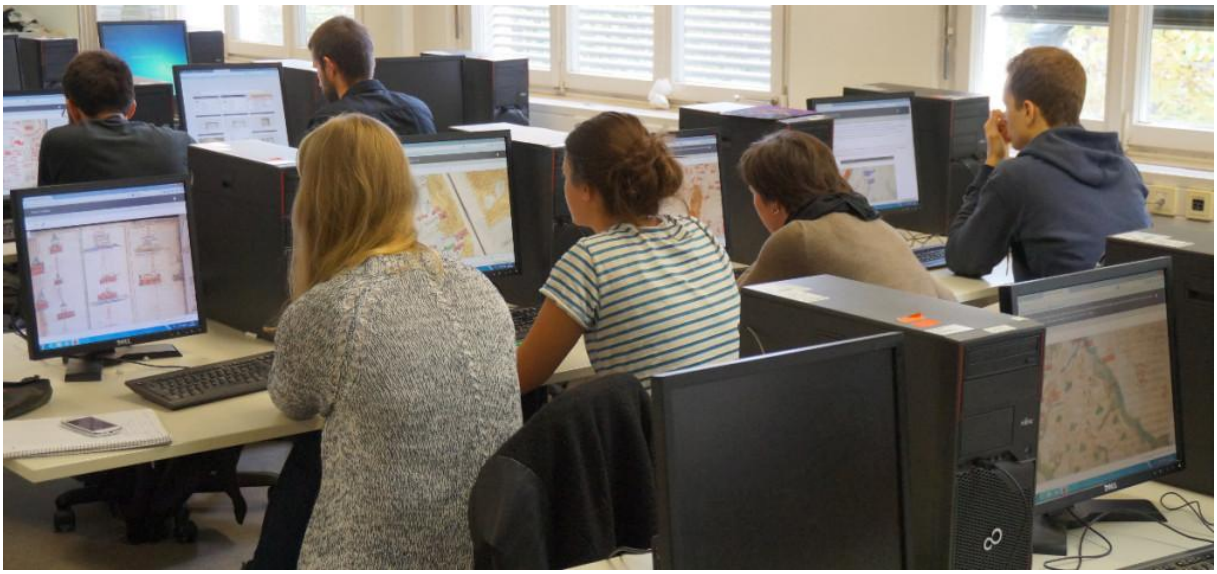


Fig.2 Impressions from the SEA CHANGE Heidelberg workshop: participants working on medieval maps.

Participants seemed to genuinely enjoy the process. Not only did we get positive feedback after the session, but several participants also followed our invitation to get permanent *Recogito* logins so that they can continue contributing after the workshop. (We recorded a further 1,648 contributions on Saturday, November 1st, the day after workshop.)

It was interesting for us to see such a clear division in terms of how the number of contributions was distributed over different task types. On the one hand, they reflect how different phases of the annotation workflow are more or less time consuming. Demarcating a place name in a text is usually a matter of a double click, for instance, whereas on a map it takes longer to navigate the image and select the area (selecting is a process that involves a mouse click, drag, and another click). Hence the roughly equal number of name identifications in texts and maps, despite the fact that more people were working on maps. Transcribing takes even more time, as we might expect; as does gazetteer resolution, i.e. searching through lists of potential gazetteer search results, and picking the one that most likely corresponds to the place name in question.

3.2 Results Mainz

For the workshop in Mainz, we followed the same procedure as in Heidelberg. In response to the low number of gazetteer resolutions (and feedback we had collected about it) we decided to re-design the user interface of this

particular *Recogito* work area beforehand, in particular with regard to where UI elements were placed, and the amount of screen real estate that was dedicated to them (e.g. giving more space to the map, while search results would be organized into groups and "folded" into collapsible lists to take up less screen space). The Mainz workshop was the first live trial run for this revised interface.

At the end of the day, we recorded a total of 7,511 contributions. These consisted of approx. 2,600 place name identifications in text (roughly an identical number to our first workshop); almost 3,200 place name identifications on images (significantly more than in the first workshop); about 620 map transcriptions (slightly less than the previous 830); 544 gazetteer resolutions; and 537 other activities such as corrections, comments, and deletions.

4. Conclusion

Overall, we were extremely happy with the amount of data our participants generated in the short time, and the continuity in terms of distribution of contributions over tasks. This seemed to show that *Recogito* is reaching a level of maturity that qualifies it for "non-expert use", beyond the confines of our Pelagios project team.

It is also interesting to speculate about where some of the differences in the results may have come from: for example, it was interesting to see significantly more place name identifications on maps in the second workshop. We assume this was simply a result of the different material. The medieval nautical charts we prepared for the second workshop are very "dense" in place names, and the place names are typically arranged in sequence, in the same orientation. So there is less need for users to search and navigate the map. That may have allowed for slightly speedier tagging. On the other hand, though, the style of lettering in these maps was rather different from last time and much more challenging for the non-expert to decipher. This may well be the reason why the number of transcriptions was lower. Furthermore, we were particularly happy to see the almost 4 times increase in gazetteer resolutions, which is an indication of the positive impact our user interface redesign had.

The two workshops were our first significant attempt at reaching out to a broader community. The results have encouraged us to look more closely into "community-sourcing" as a future strategy for Pelagios and beyond, and to evolve our approach and toolset further into this direction. However, more work and experimentation will be needed to understand factors that influence crucial aspects such as ease of use, data quality issues, and what makes the annotation process motivating and fun (in particular to users that lack expert knowledge about ancient sources and historical background). In terms of the latter, light-hearted competition clearly played a part (which we helped foster with a live feed of statistics throughout the sessions). But motivation needs more than just point scoring: one specific feedback we took away from SEA CHANGE in this regard was that people seemed to enjoy the process most when they found meaning in it for themselves. One student, for example, commented on the experience of annotating an illustrated itinerary from a medieval manuscript – a document which, from a modern person's point of view, wouldn't be considered very "map-like" in appearance. She remarked that while she was annotating the document, the geographical nature of the document would progressively start to unfurl to her. As she identified places step by step, she would begin to "see it as a map".

5. Acknowledgements

The authors wish to thank the Andrew W. Mellon Foundation, the DM2E project and the Open Humanities Awards for funding this work. Furthermore, we are indebted to our workshop hosts: Lukas Loos and Armin Volkmann from the University of Heidelberg, and Kai-Christian Bruhn and his team from the University of Applied Sciences Mainz. Last but not least we want to thank everyone who attended our workshops for their participation and enthusiasm.

6. References

[1] Elliott, T. and Gillies, S. 2009. Digital Geography and Classics. In *Digital Humanities Quarterly*. Vol 3. Number 1. http://www.digitalhumanities.org/dhq/vol/3/1/000031.html (last visited January 22, 2015)

Early Modern European Peace Treaties Online
Final Report

Dr.-Ing. Michael Piotrowski
Leibniz Institute of European History
Mainz, Germany
piotrowski@ieg-mainz.de

March 6, 2015

Abstract

Europäische Friedensverträge der Vormoderne online ("Early Modern European Peace Treaties Online") is a comprehensive collection of about 1,800 bilateral and multilateral European peace treaties from the period of 1450 to 1789, published as an open access resource by the Leibniz Institute of European History (IEG). The goal of the project funded by the Open Humanities Award was to publish the treaties metadata as Linked Open Data, and to evaluate the use of nanopublications as a representation format for humanities data.

This report describes the background of the project, the methods and tools used, the outcome, and future work.

1 Motivation

The use of databases in historical research is not new. Historical scholars have long used database management systems to store, organize, and query data they have gathered about sources, persons, places, or other items pertinent to their research questions. In many cases, these databases are never published, but even if they are made available, they tend to remain "solitary monoliths" unconnected to other data sources. There are a number of factors that are likely to contribute to situation; the following list is probably not exhaustive:

- The prevailing research culture in the humanities still awards hardly any recognition for outputs other than monographs and journal articles. In many historical research project, the primary output are thus printed publications, and no resources are allocated for preparing data for publication, let alone integrating the data produced in the project with other data sources.
- There is a general lack of (technical) coordination and a lack of standards for data and metadata (e.g., controlled vocabularies) in the humanities, which results in poor interoperability between different data sets.
- Many databases in historical research are not planned for in advance but start out as personal tools to address the specific needs of an individual scholar. Thus, in many cases "desktop" DBMS such as Filemaker or Microsoft Access are used, which tend not to scale well, and which cannot be used as a backend for a Web frontend.
- Even when a Web interface is available, the data remains isolated because typically no API is available to query the data in other ways than those offered by the human-oriented Web interface.

So, even though the individual databases are useful resources, their *full potential* is often not be realized. A case in point is the database "Europäische Friedensverträge der Vormoderne – online" ("Early Modern European Peace Treaties Online"), a comprehensive collection of about 1,800 bilateral and multilateral European peace treaties from the period of 1450 to 1789, published as an open-access resource by the Leibniz Institute of European History (IEG) in Mainz, Germany.[1] This database was created from 2005 to 2010 in a DFG-funded project with the same name.

Peace treaties between dynasties and states form an important part of our European cultural heritage. They are essential for research into early modern peacekeeping and diplomacy. "Europäische Friedensverträge der Vormoderne online" bundles manuscripts that are scattered over archives all over Europe, often hard to access, and partly undocumented. The manuscripts—in most cases the originals signed by the negotiators representing the involved powers—were digitized between 2005 and 2010 in a DFG-funded research project. All facsimiles are annotated with basic metadata, and some particularly important treaties are also available as full-text critical editions. This unique combination of digital facsimiles and critical editions has turned out to work as a well-received starting point for scholarly research in this area.

The collection data is currently stored in a relational database with a Web front-end and is one of the most popular digital offerings of the IEG. However, it has also has some shortcomings. The database is an open-access resource, but it is not machine-processable and reusable. It also lacks some important pieces of information, in particular the language or languages of the treaty texts, and the names of the undersigned negotiators. This data was collected in a later BMBF-funded project entitled "Übersetzungsleistungen von Diplomatie und Medien im vormodernen Friedensprozess. Europa 1450–1789"[2] ("Acts of translation by diplomacy and media in pre-modern peace processes. Europe 1450–1789"), which ran from June 2009 to May 2012. Researchers at the University of Augsburg gathered all the negotiators occuring in the treaties contained in the database, as well as the languages in which they are written. However, according to Penzholz and Schmidt-Rösler (2014), it was not possible to add this data to the database of treaties; instead, the scholars at the University of Augsburg created a *separate* Microsoft Access database.

The Access database is not publically available, but excerpts of the content are published as lists on a Web site[3]. Thus, even though they are based on the *same* collection of peace treaties, there exists no machine-processable link between these two databases.

Finally there is another, more conceptual problem. It is not specific to theses databases, but applies to most databases in historical research (and in many other humanities disciplines): Conventional databases are not designed to handle uncertain and contradictory data, and there is no easy way to associate certainty and provenance information with individual items. Databases in historical research thus create—usually unintentionally—an illusion of historical factuality, when, in many cases, the historical data is uncertain, scholars' interpretations of it significantly varies, and provenance information would be needed to assess its reliability.

2 Approach

Considering the issues outlined above, we conclude that conventional databases are not well suited to the requirements of historical data and research. The goal of the project funded by the Open Humanities Award was to demonstrate Linked Open Data (LOD) as a better alternative, by bringing Early Modern Peace Treaties Online to the "Linked Data cloud," allowing researchers not only to search and browse

[1] `http://www.ieg-friedensvertraege.de/`

[2] `http://www.uebersetzungsleistungen.de/`

[3] `https://www.uni-augsburg.de/de/institute/iek/projekte/historische-friedensforschung/Materialien/`

the collection but also to use and reuse the data in novel ways and to integrate it with other collections, including Europeana. By publishing our collection of European peace treaties as Linked Open Data we also wanted to make more content and data openly available for researchers to use, and make it possible to link it to other relevant information, e.g., persons and places via GND/VIAF.

In order to address the issues of uncertainty and provenance, we wanted to explore *nanopublications* as a novel approach. Nanopublications (Groth et al., 2010) were originally developed in the biomedical domain for integrating different ontologies in a common framework in order to describe scientific statements together with their context and their provenance, so that central scientific results can be unambiguously referenced and connected to their authors, and to support discovery and automatic aggregation and analysis. Nanopublications are encoded in RDF and use named graphs for grouping all information relevant for a scientific result in a single container; thus, they are compatible with the Linked Open Data approach.

Despite their highly interesting properties, the use of nanopublications in the humanities has so far only be attempted by Heßbrüggen-Walter (2013), who has used them to attribute philosophical statements—documented in their writings—to early modern philosophers. As this aspect of the project was highly experimental, we decided to proceed stepwise and first do a straightforward conversion of the existing database into RDF to make it available as Linked Open Data, and to then examine the use of nanopublications as a format for representing information about provenance and certainty in the future.

2.1 Converting Early Modern Peace Treaties Online to Linked Open Data

The process for converting the content of the existing database into LOD basically consisted of four steps:

1. **Analyzing the data.** No documentation was available for the existing database. It consists of 11 tables and numerous fields. Some of the fields have telling names, but not all of them. Another question was what the fields would actually contain. We found out that sometimes creative solutions were used. For example, the parties of a treaty are stored in a field declared as follows:

   ```
   ‘partners‘ varchar(255) NOT NULL DEFAULT ’’
   ```

   Thus, `partners` is a string field, but it does not contain the names of the parties to the treaty, but rather their IDs, e.g., `37,46,253` in string form.

   In order to determin the names of the partners, one has to first split the string, and then can look up the names in another table to find out that 37 is France, 46 is Genoa, and 253 is Naples–Sicily. This approach was used as a workaround for the problem of storing lists of variable length, which is quite tedious in a relational database. While this approach is better than hardcoding the names of the partners in every record, it moves a part of the semantics into the application, which has to know that some string fields actually contain lists of keys for a table.

   While this example is not particularly complicated, it illustrates that a thorough analysis of the database was necessary in order to accurately extract and convert the data it contains.

   During our analysis of the database, we also discovered that some potentially interesting information is only available as unstructured text, in particular references to contemporary prints and to secondary literature. We decided to skip these fields for the time being. A closer examination will be necessary to determine what additional information could realistically be extracted, i.e., with reasonable manual effort.

2. **Identifying and selecting pertinent ontologies.** We did not want to re-invent the wheel but rather build upon existing and proven ontologies for describing treaties.

3. **Modelling the information in RDF.** Once we knew how to conceptually model the information, we needed to define how to actually represent the information on a treaty in RDF.

4. **Generating the data.** Finally, we iterated over the database, extracted the information, combined it into RDF statements, and output them in a form suitable for importing them into a triple store.

At this point, we had converted the the structured metadata from the legacy database into RDF. As we expected, the conversion required a fair bit of interpretation and cleanup work, but all in all, it worked quite well.

As the basis for our data model we have, not surprisingly, used the DM2E model. Currently we have three main classes of entities, namely the *treaties*, the *treaty partners* (or signatories–but we prefer the term *partner* to avoid confusion with the negotiators, i.e., the persons who actually signed the treaties), and finally, the *locations* where the treaties were signed. We use `dm2e:Manuscript` as class for the treaties, `edm:Agent` as class for the partners, and `edm:Place` as class for the locations. Furthermore we use the following properties:

- `dc:title` for the treaty titles,
- `dc:date` for the treaty date,
- `edm:happenedAt` for linking to the location,
- `rdfs:label` for the names of partners and locations, and
- `skos:narrower` and `skos:broader` for modeling the hierarchy of partners.

The last point may need some explanation. Partners may be in a hierarchical relationship to each other to model that a power may be part of a larger entity. For example, Austria was a part of the Holy Roman Empire, whereas Milan, Mantova, and Sardinia were (at various points in time) parts of Austria. However, historical realities tend to be quite messy, so these relations are not necessarily "part-of" relations in the strict sense; for example, Austria also had territories outside the Empire. The hierarchy also contains "fictitious partners" as a help for searching; for example, introducing *Switzerland* or *Parts of the Empire* as "fictitious partners" makes it easier to search for treaties concerning certain regions of Europe. This pragmatic approach was taken over from the legacy database, as we think it makes sense, at least for the time being.

To link the treaties to the treaty partners we used the `dc:contributor` property. This usage stretches the meaning of "contributor" a bit, but we only use it as a provisional solution, as we will reconsider the modeling when moving to nanopublications.

If we consider a specific treaty, such as the *Provisional convention of subsidy* between Great Britain, the States General, and Austria, we have the following data:

Property	Value
Type	dm2e:Manuscript
Title (`dc:title`)	Provisorischer Subsidienvertrag (de)
Date (`dc:date`)	1746-08-31
Contributor (`dc:contributor`)	Austria, Great Britain, States General
Happened at (`edm:happenedAt`)	The Hague

This display is somewhat simplified for illustration. For reference, figure 1 shows the last page of the treaty; the last sentence before the seals and signatures gives the place and the date: "Fait à La Haye le trente un du Mois d'Aout de l'année mille Sept cent quarante Six."

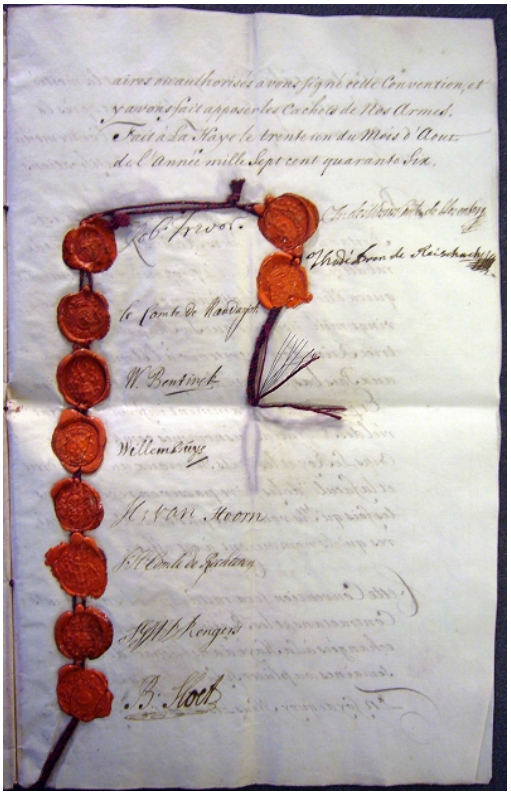


Figure 1: Provisional convention of subsidy between Great Britain, the States General, and Austria (Nationaal Archief, Den Haag, Staten-Generaal, nummer toegang 1.01.02, inventarisnummer 12597.187)

We loaded the data into Fuseki and set up a server at `http://data.ieg-friedensvertraege.de/`. Since one cannot really "see" Linked Open Data, we also set up Pubby, a Linked Data frontend for SPARQL endpoints, which gives the data a friendlier face. For example, the screenshot below shows how the information on the Friedenspräliminarien von Breslau (in English known as Treaty of Breslau) is presented in Pubby.

As noted above, there is not really much to "see" about the data, so there is not really much to show; what is exciting is the potential it has for automatic processing. As a low-key example, the homepage at `http://data.ieg-friedensvertraege.de/` currently shows a "live" list of all treaty partners, i.e., when one loads the Web page, a query is sent to the SPARQL endpoint to retrieve all entities of type `edm:Agent` (see figure 3). We intend to replace this list with a more interesting example such as a map showing the treaty locations, which could look like the mockup shown in figure 4.

However, in order to be able to draw such a map, the geographical coordinates of the treaty locations must obviously be known, whereas the original database only contains placenames. This requires that they are linked to a suitable data source. This step is necessary in any case in order to make the data not just open but also actually linked.

Since there are 478 locations and 201 partners, we used an automated process to look up the names used in the database in the GND. As expected, many of the names are ambigous, whereas others are not found at all. Here are some examples, illustrating the variation with respect to GND IDs found for some treaty locations:

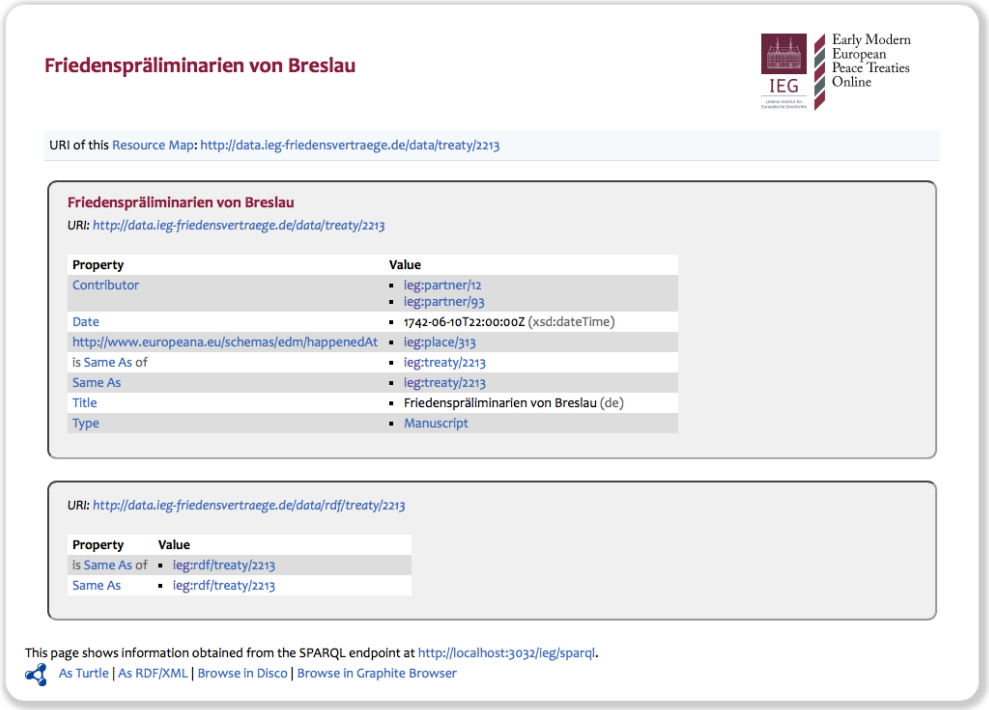


Figure 2: Screenshot of Pubby

ID	Name	GND
33	Altranstädt	4079738-7
4	Hubertusburg	5119515-X, 5119512-4
91	Malmö	4114951-8
6	Tyrnau	10172490-1, 500513-9, 10179031-4, 4555737-8, 7582117-5, 4696473-3, 1044374594, 4078490-3, 4696475-7

It is clear that such ambiguities cannot be resolved automatically, in particular, there is no guarantee that the correct entity is in fact among those found. We did not have resources in the project to perform manual resolution of location and partner links. However, it is one of the advantages of nanopublications that data can be qualified with provenance information, so that, for example, data added by automatic processes is given a lower certainty than data added by a human expert. The next section talks about nanopublications in some more detail.

2.2 Peace Treaties as Nanopublications

The second main goal of this project was to to explore the application of this approach to research in the humanities and to represent the key metadata about peace treaties (date, place, signatories, powers, type of treaty, etc.) as nanopublications. One important advantage of nanopublications is that they allow for associating provenance information with claims, which, in turn, also helps dealing with uncertain, unconfirmed, or conflicting claims.

Figure 5 shows an example of a nanopublication. It states—in the *assertion* part—that that treaty

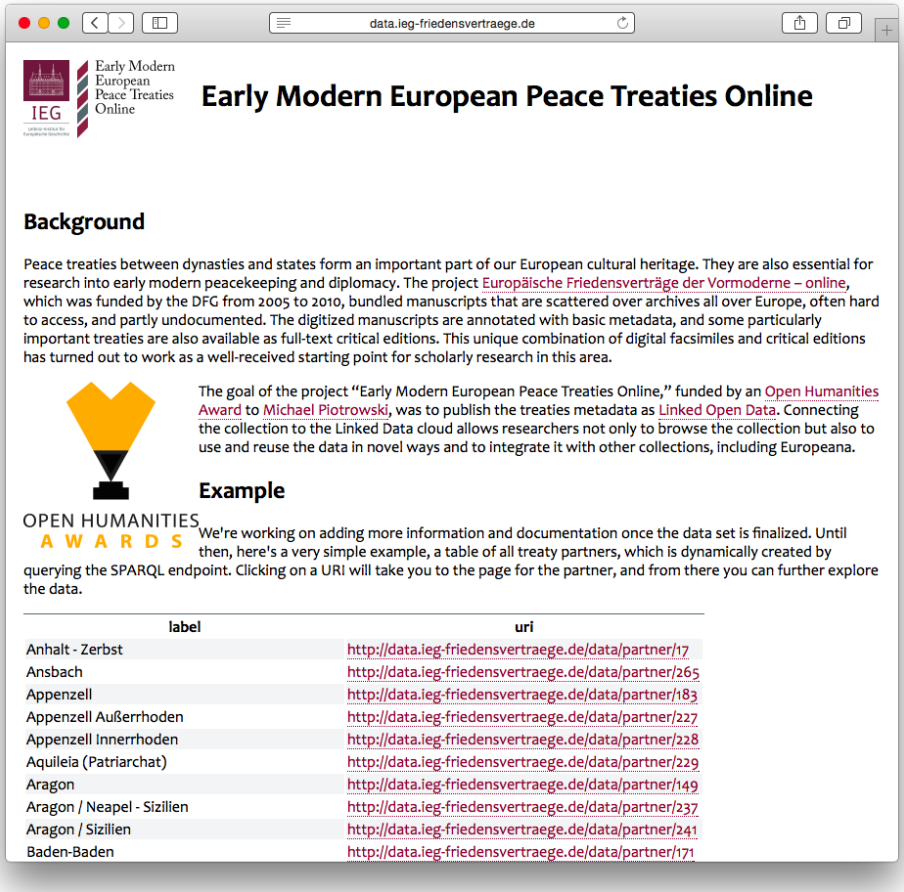


Figure 3: Screenshot of the `http://data.ieg-friedensvertraege.de/` home page

partner 12 ("Austria") is identical with the entity with the GND ID 10105216-9. The *provenance* part documents the provenance of this assertion; in this case, it was automaticall created by a script called *autolinks*, which automatically adds potential links to the GND. The *pubinfo* part, finally, contains metadata about the nanopublication as a whole.

As it is known that the assertion was generated automatically, it can be treated with the appropriate caution. In fact, while the GND ID given here does refer to an entity called "Austria," it is actually the British zone of occupied Austria, 1944–1955, which clearly cannot be the correct reference in the context of our early modern peace treaties.[4] This is, however, not a problem: First, as the provenance information is given, it is possible to filter data on the basis of this information; second, the nanopublications approach makes it possible to explicitly refute this assertion by another nanopublication.

The Nanobrowser by Kuhn et al. (2013) implements a user interface to nanopublications that specifically supports this type of interactions, so that a researcher can easily reject or support assertions. We are in close contact with the author of the Nanobrowser (which is open-source software) and are working on adapting it to our needs.

[4] A more likely reference would be either 4043271-3 "Austria" or 4075601-4 "Archduchy of Austria."

Figure 4: Mockup of a map display for treaty locations

```
@prefix nanopub: <http://www.nanopub.org/nschema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prov:    <http://www.w3.org/ns/prov#> .
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd:     <http://www.w3.org/2001/XMLSchema#> .
@prefix owl:     <http://www.w3.org/2002/07/owl#> .
@prefix :
 <http://data.ieg-friedensvertraege.de/nanopub/> .

:/1/head {
  :/1 a nanopub:Nanopublication ;
    nanopub:hasAssertion       :/1/assertion ;
    nanopub:hasProvenance      :/1/provenance ;
    nanopub:hasPublicationInfo :/1/pubinfo .
}

:/1/assertion {
  <http://data.ieg-friedensvertraege.de/data/partner/12>
  owl:sameAs <http://d-nb.info/gnd/10105216-9> .
}

:/1/provenance {
  :/1/assertion prov:wasGeneratedBy :pfeffer/autolinks/gnd/v0.1 .
}

:/1/pubinfo {
  :/1 dcterms:creator <http://d-nb.info/gnd/102572836X> ;
    dcterms:created "2015-01-28T11:32:30.758274Z"^^xsd:dateTime ;
    dcterms:rights <https://creativecommons.org/publicdomain/zero/1.0/> ;
    dcterms:rightsHolder <http://www.ieg-mainz.de/> .
}
```


Figure 5: Example nanopublication

Due to problems finding qualified personnel at the outset of the project, we lost about a month, and with the holiday season before the end of the project we lost more time. We thus could not complete the work on nanopublications during the allotted time. However, we have laid important foundations, and we continue to pursue this line of research even after the end of the project.

3 Conclusion and Outlook

In the project funded by the Open Humanities Award, we have converted the metadata of Early Modern Peace Treaties Online from a relational database into RDF and made it available as Linked Open Data. While we could not finish our work on nanopublications in the time frame of the project, we were able to lay the groundwork, and we are currently setting up the Nanobrowser and the tools for generating nanopublications from the RDF triples that we produced in the project.

For the work described here, I contracted with Prof. Magnus Pfeffer of the Stuttgart Media University (HDM). We are continuing the work on nanopublications as an unfunded research project. Prof. Dr. Kai Eckert, who previously worked in DM2E, and who has now also joined HDM, will now team up with us and contribute valuable experience from DM2E. We are currently working on a joint peer-reviewed paper, which will also cover the use of nanopublications.

References

Groth, Paul, Andrew Gibson, and Jan Velterop (2010). The anatomy of a nanopublication. *Information Services and Use*, 30(1–2):51–56. doi:10.3233/ISU-2010-0613.

Heßbrüggen-Walter, Stefan (2013). Tatsachen im semantischen Web: Nanopublikationen in den digitalen Geisteswissenschaften? In Peter Haber and Eva Pfanzelter, eds., *Historyblogosphere*. München: Oldenbourg. doi:10.1524/9783486755732.149.

Kuhn, Tobias, Paolo E. Barbano, Mate L. Nagy, and Michael Krauthammer (2013). Broadening the scope of nanopublications. In *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013)*, vol. 7882 of *Lecture Notes in Computer Science*, pages 487–501. Berlin/Heidelberg: Springer. doi:10.1007/978-3-642-38288-8_33.

Penzholz, German and Andrea Schmidt-Rösler (2014). Die Sprachen des Friedens – eine statistische Annäherung. In Johannes Burkhardt, Kay Peter Jankrift, and Wolfgang E. J. Weber, eds., *Sprache. Macht. Frieden*, pages 311–322. Augsburg: Wißner. URL `http://www.uni-augsburg.de/institute/iek/scripts/Sprache_Macht_Frieden_Penzholz-Schmidt-Roesler.pdf`.

The research described in this report was funded by the DM2E project as part of the Open Humanities Awards.

Wittgensteins Nachlass: Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind

Max Hadersbeck, Alois Pichler, Florian Fink, Daniel Bruder, Ina Arends, Johannes Baiter
Maximilian.Hadersbeck@lmu.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München,
Wittgenstein Archives at the University of Bergen (WAB).

1 EINLEITUNG

In dem Vortrag berichten wir über Erfahrungen, Erkenntnisse und Erweiterungen unserer schon seit 2 Jahren im Einsatz befindlichen FinderApp WiTTFind, die mit Hilfe von computerlinguistischen Verfahren den Open Access zugänglichen Teil des Nachlasses von Ludwig Wittgenstein (Wittgenstein Source, 2009) nach Wörtern, Phrasen, Sätzen und semantischen Begriffen im „Zusammenhang des Satzes“[1] durchsucht.

Im Sommer 2014 gewannen wir mit WiTTFind den EU-AWARD, der vom EU-Projekt Digitised Manuscripts to Europeana (DM2E) ausgeschrieben wurde, verbunden mit der expliziten Aufforderung zur Öffnung unseres Finders für andere Projekte der Digital Humanities. Darauf hin entwarfen wir in der disziplinübergreifenden Wittgenstein Sommerschule am CIS im Juni 2014 und in Diskussionen mit Fachleuten der Philosophie und Digital Humanities Verbesserungsmöglichkeiten, die mittlerweile in der neuen Version implementiert sind. Die Weboberfläche unseres Finders wurde optimiert, („rich-client“), jetzt können mehrere Dokumente parallel durchsucht werden, eine lemmatisierte symmetrische Vorschlagssuche und ein Faksimile E-Reader sind integriert. Der Faksimile E-Reader erlaubt es nun, dass die Faksimiles der Edition durchblättert und gefundene Textstellen automatisch visuell hervorgehoben werden. Neben den Weiterentwicklungen der FinderApp setzten die Wittgensteinforscher unseren Finder für semantische Untersuchungen ein und gewannen aus dieser Arbeit wichtige Erkenntnisse z.B. zum Thema des Verstehens in Wittgensteins Big Typescript.[2]

Der wichtigste Mehrwert unseres Finders besteht allerdings darin, dass wir die vom EU-AWARD geforderte Öffnung unseres Finders für andere Projekt konsequent umsetzten. Für die Texte der Edition, die unser Finder durchsucht, gibt es eine XML-TEI P5 kompatible Document Type Definition (DTD). Die Programme, Faksimile E-Reader und Tools sind unter der Bezeichnung „Wittgenstein Advanced Search Tools“ (WAST) in einem „docker“-Softwarecontainer zusammengefasst und werden „open source“ verfügbar sein. Somit ist unsere FinderApp mit ihren WAST-Tools in anderen Projekten der Digital Humanities einsetzbar.

Die folgende Abbildung zeigt eine Suchanfrage an unseren Finder WiTTFind:

http://wittfind.cis.uni-muenchen.de:

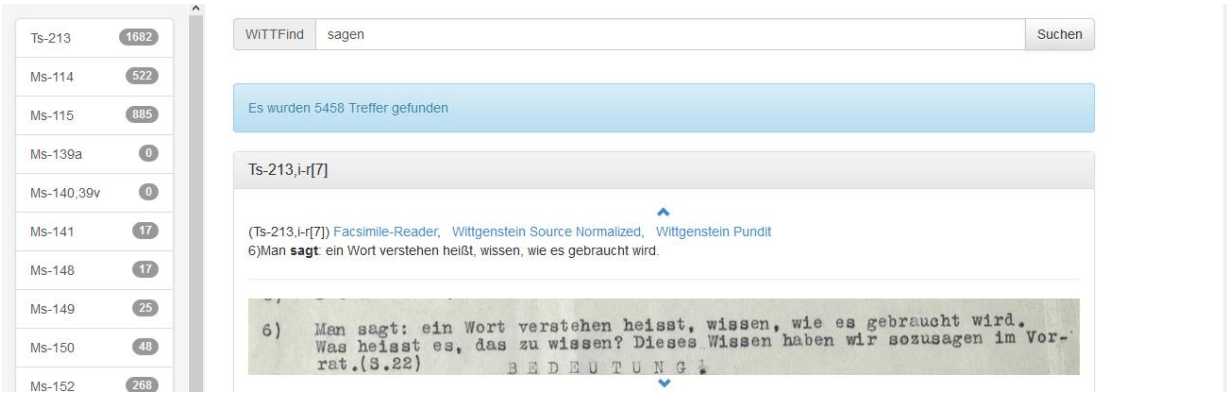


Bild 1: Suchanfrage bei WiTTFind

[1] http://www.wittgensteinsource.org/Ts-213,1r[4]_n

[2] http://www.wittgensteinsource.org/Ts-213_n

2 Erkenntnisse aus der Zusammenarbeit Computerlinguistik und Philosophie

2.1 Verbesserte Benutzeroberfläche unseres Finders

Eine der ersten Erkenntnisse unserer Zusammenarbeit war, dass die Benutzeroberfläche unserer FinderApp auf die Bedürfnisse der jeweiligen Forschergruppe abgestimmt sein muss: die Forscher sollen sich auf der Webseite „wiederfinden“. Nur dann ist die Einstiegshürde nicht zu hoch, und die Bereitschaft mit dem Finder zu arbeiten steigt. Erst für fortgeschrittene Benutzer werden in einer tieferen Schicht globale Einstellungsmenus sichtbar und spezielle Parameter einstellbar. Als Kompromiss zwischen Komplexität und gewohnter Suchmaschinenarbeit können die Nutzer verschiedene Suchumgebungen auswählen (siehe Bild 1): „Regelbasiertes Finden“, „Semantisches Finden“, „Graphisches Finden“, „Statistische Suche“ und „Geheimschriftübersetzer“.

Bild 2: Suchumgebungen bei WiTTFind

Damit die zahlreichen Suchmöglichkeiten bei WiTTFind auf einen Blick sichtbar sind, programmierten wir fachspezifische Hilfeseiten mit Beispielen:

Bild 3: Hilfeseiten bei WiTTFind

2.2 Video-Tutorials zur Nutzung von WiTTFind

Zum erleichterten Einstieg bei WiTTFind gibt es jetzt zwei Video-Tutorials in deutscher und englischer Sprache unter folgendem Link:

http://wittfind.cis.uni-muenchen.de/tutorial

2.3 E-Reader für die Faksimile

Gerade bei komplexen Editionen mit vielen handschriftlichen Einfügungen und Streichungen, wie der des Nachlasses von Ludwig Wittgenstein, ist es für die Editionswissenschaftler eine Herausforderung, den Editionstext in der niedergeschriebenen Form als HTML-Text in einem Browser darzustellen. In der neuen Version unserer FinderApp programmierten wir einen eigenen Faksimile E-Reader, der es erlaubt, komplementär durch die Faksimile der Edition zu blättern und gleichzeitig die gefundenen Textstellen im Bild hervorhebt.

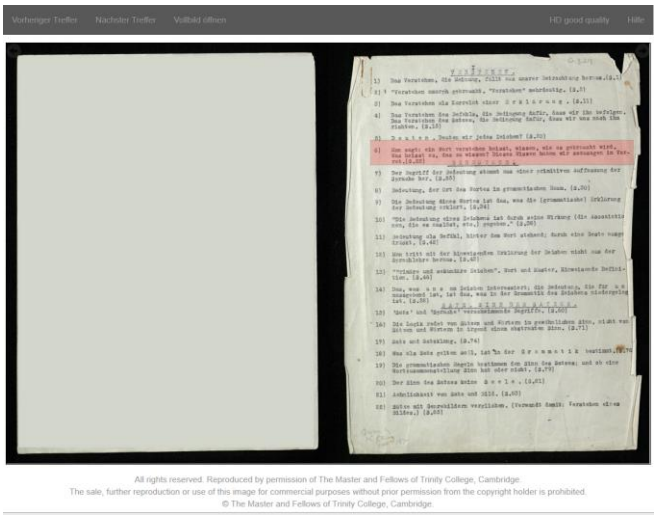


Bild 4: Faksimile Reader bei WiTTFind

2.4 Lemmatisierte Vorschlagssuche mit statistischen Angaben

Die Arbeit mit WiTTFind zeigte, dass eine komfortable Vorschlagssuche, die den gesamten Wortindex der Edition mit Frequenzlisten im Hintergrund hält, einen sehr guten Einstieg in die eigentliche Suche darstellt. Hierhin zielt unsere neueste Erweiterung von WiTTFind, eine komfortable Index-Suchfunktion, die auf einen symmetrischen Suchindex basiert. Dieser Index greift auf Einträge des zugrunde liegenden Lexikons und Wort-Frequenzlisten der Texte der Edition zurück. Dem Anwender werden nach Eingabe von wenigen Buchstaben alle Wörter mit der Häufigkeit des Auftretens im Text automatisch aufgezeigt, in denen die eingegebenen Buchstaben vorkommen; dazu werden auch noch die morphologischen Varianten dieser Wörter angezeigt. Diese Art der Autovervollständigung ist eine völlig

neue Technologie, da bisherige Autovervollständigungen die eingegebenen Buchstaben nur um die Wörter ergänzen, die mit diesen Buchstaben beginnen.

3 Von Daten zu Erkenntnissen

3.1 Semantisches Suchen: Wortfelder

Ein großes Problem semantischer Untersuchungen mit Wortfeldern stellt die Disambiguierung der Wortfeldbegriffe dar. Mit Hilfe unseres elektronischen Lexikons, der syntaktischen und semantischen Disambiguierung über Part of Speech Tagging und lokale Grammatiken können neben Einzelwörter auch Wortphrasen einem Wortfeld zugeordnet und disambiguiert werden.
Ein einfaches Beispiel wurde um das semantische Feld von "Verstehen" ausgearbeitet. Welches Interesse an Verstehen hat Wittgenstein im Big Typescript? Eine Suche nach *<N> verstehen* [Substantiv + „verstehen“] im Big Typescript ergibt, dass dort ganz klar das Verstehen von Wörtern, Sätzen, Sprachen, Befehlen ... allgemein: das Verstehen von sprachlichen Zeichen, im Vordergrund steht. Daneben gibt es aber auch bereits eine gewisse Aufmerksamkeit auf das Verstehen von Menschen und Menschlichem: von Handlungen, Gebärden, Gesten. Diese Aufmerksamkeit nimmt in Wittgensteins Spätwerk beständig zu, was eine Suche nach *<HUM> verstehen* [Substantiv für Menschliches + „verstehen“] bestätigt.
Ein zweites, komplexeres Beispiel wurde um das semantische Feld von "Grammatik" ausgearbeitet. Zuerst baten wir Wittgensteinexperten, uns eine Liste von 10-15 Wörtern zu geben, welche ihrer Ansicht nach im Wortfeld von "Grammatik" zentral sind. Dazu gehören z.B. "Anwendung", "Regel", "Kalkül" und "System". Daraufhin wurden diese Wörter im Lexikon über den Begriff "Grammatik" vernetzt. Eine WiTTFind-Suche nach *Grammatik* wird dann nicht nur Stellen mit "Grammatik" ergeben können, sondern auch Bemerkungen, welche eine Bündelung von Begriffen aus dem Wortfeld aufweisen. Erste Anwendungen ergaben, dass Wittgenstein im Big Typescript tatsächlich einen regelfixierten Begriff von Grammatik verfolgt, während dieser Aspekt später abgeschwächt werden wird (vgl. Szeltner 2013).

4 Synergien: Unsere FinderApp für andere Digital Humanities Projekte

4.1 Vorbemerkung

Wie vom DM2E Projekt bei der Preisverleihung gefordert, öffneten wir unsere FinderApp für andere Projekte der Digital Humanities. Editionsprojekte müssen ihre Dokumente in unser reduziertes XML-TEI P5 Format (CISWAB) konvertieren und die Open-Source Software *docker*[3] auf ihrem Rechner installieren. Dann können sie unseren Finder bei ihren Editionstexten anwenden. Zur Darstellung und Highlighting der Treffer im Faksimile sind allerdings umfangreiche OCR-Arbeiten notwendig. In den nächsten Unterkapiteln beschreiben wir im Detail, wie unser Finder einsetzbar wird.

4.2 Die Texte der Edition

Unsere FinderApp findet Wörter, semantische Begriffe und Satzphrasen über mehrere Dokumente hinweg, sofern die Dokumente in unserem XML-TEI-P5 Format vorliegen. Wir nennen dieses XML-Format CISWAB und beschreiben es in einer eigenen Document Type Definition (DTD). Die einzelnen Dokumente sind bis auf Satzebene über Siglen eindeutig zu spezifizieren:
(z.B. `<s n="Ts-213,i-r[7]_1" ana="facs:Ts-213,i-r abnr:7 satznr:15">6)Man sagt: ein Wort verstehen heißt, wissen, wie es gebraucht wird.</s>`)

4.3 Elektronisches Vollformenlexikon

Zu den Texten einer Edition benötigt unsere FinderApp ein elektronisches Lexikon im DELA Format (Laboratoire d'Automatique Documentaire et Linguistique, Paris). Bei der Erstellung des Lexikons können wir behilflich sein, da wir am CIS das größte deutsche Vollformenlexikon erstellt haben.

4.4 Syntaktische Disambiguierung: Part of Speech Tagging

Grundvoraussetzung für die syntaktische Disambiguierung ist es, dass die Texte mit einem Part of Speech Tagger bearbeitet werden. Zu unseren WAST-Tools gehört das automatische Taggen der Texte. Dazu verwenden wir den *treetagger* von Dr. Helmut Schmid, der am CIS entwickelt wird. Der *treetagger* konvertiert

[3] siehe: https://www.docker.com/

die Textdatei in eine getaggte XML Datei, die die Eingabedatei für unsere FinderApp darstellt.

4.5 Darstellung der Treffer im Faksimile Reader

Um die Treffer in unserem Faksimile-Reader darzustellen, müssen die Faksimile mit der open source Software *tesseract* bearbeitet werden, und je nach Qualität der Faksimiles manuell nachbearbeitet werden. Wir entwickelten Tools, die diese manuelle Arbeit erleichtern.

4.6 Praktische Voraussetzung zur Verwendung unserer FinderApp

Wir haben unser Ziel, dass die FinderApp WiTTFind und die WAST-Tools möglichst auf jedem Rechner lauffähig sind, erreicht. Mit Hilfe der neuesten Open Source Software Technologie *docker* werden die unterschiedlichen Programmiersprachen und Libraries, die wir einsetzen, in einem Softwarecontainer, genannt WAST-dockerimage, zusammengefasst. Jeder Anwender, der auf seinem Rechner die *docker*-Serversoftware installiert hat, kann das WAST-dockerimage herunterladen und virtualisiert läuft die FinderApp WiTTFind unter dem Dockerserver auf dem Rechner. Die Dockerserversoftware funktioniert nahezu unter jedem Betriebssystem (Linux, Windows, MACOS).

4.7 Vorstellung und Vorführung unseres Finders auf der Tagung

Neben diesem Vortrag wollen wir auf der Tagung in einem Poster den Aufbau und den Einsatz der FinderApp WiTTFind als Open Source Tool vorstellen: Die optimierte Browseroberfläche, zugrunde liegende Texte der FinderApp, Faksimile mit OCR, Faksimile Reader und den Einsatz des Finders als Open Source Programm. Für Interessierte wird die FinderApp unter verschiedenen Betriebssystemen an Laptops vorgeführt.

5 EU-AWARD und Publikationen

EU AWARD 2014: http://dm2e.eu/open-humanities-awards-round-2-winners-announced/

Max Hadersbeck, Alois Pichler, Florian Fink, Øyvind Liland Gjesdal: Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST). Digital Access to Textual Cultural Heritage 2014 (DaTeCH 2014) Madrid: 91-96

Szeltner, Sarah: 'Grammar' in the Brown Book. Papers of the 36th International Ludwig Wittgenstein-Symposium, vol 21. Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society; 2013.

Wittgenstein Source: Bergen Text and Facsimile Edition. In: Pichler A., collaboration with, Krüger H.W., Lindebjerg A., Smith D.C.P., BruvikT.M., Olstad V., editors. Bergen: Wittgenstein Archives at the University of Bergen; 2009. http://www.wittgensteinsource.org/