Historical Text Reuse Data Workshop
July 12-13, 2017

Final Report
(October 2017)

Universität Leipzig
Augustusplatz 10-11 – 04109 Leipzig

Within the framework of the BMBF funded *Global Philology Planning Project*, Dr. Monica Berti from the University of Leipzig has organized a workshop on *Historical Text Reuse Data*. For information on the organizer, please visit http://www.monicaberti.com.

The workshop was held in the Institute of Computer Science at the University of Leipzig on July 12-13, 2017. The program of the workshop is available through this link: http://www.dh.uni-leipzig.de/wo/historical-text-reuse-data-workshop/. Attendees and speakers twitted the event with the hashtag **#GPTextReuse2017**. Tweets are available thorugh this links: https://twitter.com/hashtag/gptextreuse2017?f=tweets&vertical=default&src=hash

The *Historical Text Reuse Data* workshop was focused on automated and manual methods for detecting and annotating text reuse of historical languages in a digital environment. The goal was to present and explore tools and techniques for automatic detection of text reuse, and also data models for annotating text reuse elements deriving from both automatic and manual work.

Considering the nature of text reuse detection and annotation, the workshop was also focused on textual and translation alignment techniques for comparing reuses in different texts and different languages. The discussion was also devoted to named entities recognition in order to explore techniques for identifying named entities pertaining to text reuse (e.g., author names and work titles) and therefore creating catalogs of reused authors and works.

22 speakers were invited to participate in the workshop. The speakers belonged to institutions of 8 different countries (Austria, Belgium, France, Germany, Great Britain, Italy, the Netherlands and the United States of America) and their nationalities represented 9 languages (Arabic, Dutch, English, Farsi, French, German, Greek, Italian, and Japanese).

The two days of the workshop had two sessions each (one in the morning and one in the afternoon) with a total of 12 presentations on the first day and 10 presentations on the second day.

July 12, 2017 - Morning Session
After welcoming speakers and attendees and after the introduction to the workshop by **Monica Berti** and **Gregory R. Crane**, the first part of the morning was devoted to the presentation of projects and research problems pertaining to traditional questions of text reuse.

Stefan Schorn from the Catholic University of Leuven and **Erle Monfis** from Brill Publishers in the Netherlands presented issues and perspectives of *Die Fragmente der Griechischen Historiker* project, which is the biggest collection of ancient Greek historical quotations and text reuses. Stefan Schorn focused on questions concerning the history and the arrangement of the collection from a philological and historiographical point of view, while Erle Monfis presented characteristics and future developments of the online version of the collection (http://referenceworks.brillonline.com/cluster/Jacoby%20Online).
After these two presentations, **Monica Berti** focused on models and solutions for annotating and citing historical text reuse data in a digital environment. Examples and test cases were drawn from her two ongoing projects: the *Digital Fragmenta Historicorum Graecorum* (http://www.dfhg-project.org) and the *Digital Athenaeus* (http://digitalathenaeus.org).

The second part of the morning had three speakers focusing on different aspects of the workshop: text reuse detection and classification, and automatic references to text reuses.
Stylianos Chronopoulos from the University of Freiburg presented his work on detecting and classifying text reuse in the *Onomasticon* of the ancient Greek grammarian Julius Pollux.
David A. Smith from Northeastern University in the USA presented specific and technical aspects of text reuse techniques for iterative corpus construction.
The last paper of the morning was presented by **Matteo Romanello** from the Deutsches Archäologisches Institut in Berlin, who showed the result of his PhD work on automatic extractions of references to text reuses. In particular the presentation was focused on references to still preserved historical texts and to historical texts that are now lost and preserved only through quotations and text reuses (canonical vs. fragmentary texts).

July 12, 2017 - Afternoon Session
Given that **Roland Wittwer** from the Berlin-Brandenburgische Akademie der Wissenschaften was not able to join the workshop and present his work on the collection of text reuses of ancient Greek physicians (*Corpus Medicorum Graecorum* - http://galen.bbaw.de), the afternoon session had 4 presentations and the final discussion of the day. The topics included automatic text reuse detection of historical texts and the implementation of platforms and tools for text reuse annotations.

Richard Eckart de Castilho from the Technische Universität in Darmstadt presented his work on developing and implementing *WebAnno*, which is a web-based platform for many different kinds of manual and automatic annotations of texts (https://webanno.github.io). The speaker presented also the new project INCEpTION funded by the Deutsche Forschungsgemeinshaft (https://www.ukp.tu-darmstadt.de/research/current-projects/inception/).
Marie Revellio from the University of Konstanz presented her work on computational methods of text reuse detection focusing on examples from late antique prose.

During lunch break, Monica Berti and Gregory R. Crane had a long meeting with Richard Eckart de Castillho in order to plan future collaborations between the team of *WebAnno* and the *Global Philology Project*. The goal is to expand collections of historical text reuse data using *WebAnno* and to help implement the *WebAnno* platform with many new use cases.

The second part of the afternoon had two presentations focusing on text reuse in two different languages.
Sarah Bowen Savant from the Aga Khan University of London presented text reuse in Arabic sources, while **Alexandra Trachsel** from the University of Hamburg presented her work on text reuses and quotations from the small corpus of the ancient Greek author Demetrios of Scepsis.

The **discussion** at the end of the afternoon was an opportunity to summarize the many different topics pertaining to text reuse that were presented during the day. In particular the focus was on the common problems concerning text reuse in different historical languages and the need to develop common and shared platforms and environments for collecting and disseminating data.

July 13, 2017 - Morning Session
The morning session of the second day of the workshop gathered speakers with presentations on traditional and digital problems concerning text reuse. The day was

also important for the introduction of the topic concerning textual alignment related to text reuse.

Joshua M. Smith from the Johns Hopkins University in the USA presented his database of quotations and text reuses of Homer's works in the ancient commentaries to texts.
Marianne Reboul from the Sorbonne University presented her PhD work on automatic textual alignment of French translations of the *Odyssey* of Homer from the 16th to the 20th century. This presentation introduced the topic of text reuse in different languages and the discussion was important for the speaker, who has in the meantime finished her PhD dissertation and successfully passed the final PhD defense.
Martin Potthast from the University of Weimar presented his work on *Picapica*, which is a text reuse search engine (http://www.picapica.org). Martin Potthast is now Junior Professor of Text Mining at the University of Leipzig and this presentation was an important occasion for starting a collaboration with the *Global Philology Project* for text reuse detection.

The second part of the morning was focused on translation alignment issues and textual alignment platforms.
Uta Koschmieder from the University of Halle presented her ongoing work on ancient historical text reuses in Greek, Syriac, and Armenian and on problems concerning the manual alignment of these texts. Uta Koschmieder is a PhD candidate, who started this work for her MA dissertation at the University of Halle in collaboration with the University of Leipzig under the supervision of Monica Berti.
Tariq Yousef from the University of Leipzig presented the work that he is currently implementing for an online alignment editor (http://ugarit.ialigner.com) for the *Global Philology Project*. Tariq Yousef presented also his work on the *Dynamic Lexicon* (http://dynamiclexicon.com), whose goal is to extract bilingual lexicons automatically from pre-aligned parallel texts by using information retrieval techniques.

July 13, 2017 - Afternoon Session
The last session of the workshop included different topics: manual detection and annotation of text reuse in different languages, and the development of platforms for annotating named entitities.

Takayoshi Oshima from the University of Leipzig presented his work on building a database with a collection of ancient Sumerian proverbs. Through specific examples, the speaker was able to present philological and linguistic problems concerning text reuse in Ancient Near Eastern literature.

Rainer Simon from the Austrian Institute of Technology presented new developments of *Recogito* (http://recogito.pelagios.org), which is a platform for annotating geographical information on texts and images. This tool is relevant for annotations of named entities within historical text reuses.
Chiara Palladino from the University of Leipzig and the University of Bari, who is also part of the team working on *Recogito*, presented her ongoing PhD work on collecting and annotating quotations, text reuses and fragmentary texts of ancient Greek geographers.

Maryam Foradi from the University of Leipzig is an expert of translation techniques and alignment of ancient Greek and Farsi. She presented her ongoing PhD work on translation strategies and information extraction. Maryam Foradi is also leading the *Open Persian* project at the University of Liepzig, which is an essential component of the *Global Philology Project*:
http://www.dh.uni-leipzig.de/wo/open-philology-project/open-persian/.
The last presentation was by **Jochen Tiepmar** from the University of Leipzig and from the Competence Center for Scalable Data Services and Solutions (ScaDS), who presented his ongoing PhD work on the *Canonical Text Services Protocol* (CTS): http://cts.informatik.uni-leipzig.de/Canonical_Text_Service.html. This protocol defines interaction between a client and a server providing identification of texts and retrieval of canonically cited passages of texts. This work is particularly relevant for citing and retrieving texts and their textual reuses. This protocol is also used in the *Digital Athenaeus* project of Monica Berti (http://cts.informatik.uni-leipzig.de/asusedin.html).

The **discussion** at the end of the afternoon was an opportunity to summarize the many different topics pertaining to text reuse that were presented during the two days of the workshop. In this case, the focus was on text reuse in different languages and on how to address it and on the need of developing common and shared platforms and environments for aligning texts in the same and in different languages and on disseminating the resulting data.