

Digital Fragmenta Historicorum Graecorum (DFHG)

XML

This long paper presents the *Digital Fragmenta Historicorum Graecorum* (DFHG) project (<http://www.dfhg-project.org>). The *DFHG* is the digital version of the five volumes of the *Fragmenta Historicorum Graecorum* (FHG), which is the first big collection of ancient Greek historical fragments published by Karl Müller (1841-1873). The *FHG* is a corpus of quotations and text reuses (fragmenta) of 636 ancient Greek fragmentary historians preserved by Classical sources. Fragmentary authors date from the 6th century BC through the 7th century CE and, except for the first volume, are chronologically distributed. Fragments are numbered sequentially and arranged by works and book numbers with Latin translations, commentaries, and critical notes.

The *DFHG* is not a new edition of ancient Greek fragmentary historians, but a new digital resource to provide textual, philological, and computational methods for representing fragmentary authors and works in a digital environment. The reason for choosing the *Fragmenta Historicorum Graecorum* depends not only on an interest in Greek fragmentary historiography, which provides a rich collection of complex reuses of prose texts, but also on the necessity of digitizing printed editions and preserving them as structured machine readable corpora that can be accessed for experimenting with text mining of historical languages. Moreover, the *FHG* is still fundamental to understand recent editions of Greek historical fragments and in particular *Die Fragmente der griechischen Historiker* (FGrHist) by Felix Jacoby, who spent his life to change and improve the collection created by Karl Müller. Finally, the corpus of the *FHG* is open and big enough to perform computational experiments and obtain results.

This paper presents tools and services that have been developed by the project, not only for accessing the entire collection of the *FHG*, but also for providing a new model that can be applied to other collections of fragmentary authors in order to visualize and explore their complex data and connect it with external resources for further developments. The presentation is organized according to the following topics:

- **Visualization of DFHG contents.** The *DFHG* appears as an Ajax web page automatically generated by a PHP script querying an SQL database of the entire *FHG*, which is accessible by browsing the whole collection or single volumes through a slide in/out navigation menu. The navigation menu allows scholars to navigate the *FHG* with a comprehensive and detailed view of the structure of the entire collection and to jump to the relevant section without reloading the page. This kind of visualization is very helpful because the printed version of the *FHG* doesn't contain detailed tables of contents of its volumes, but only short and sometimes incomplete lists of authors published in the collection.
- **Access to the DFHG.** The *DFHG Digger* filters the whole collection according to authors, works, work sections, and book numbers, while the *DFHG Search* function is performed on fragments, translations, commentaries and source texts. Results show the number of occurrences in each *DFHG* author and searched words are highlighted in the text. They also display, when available, the lemmatization of inflected forms and the disambiguation of named

entities through external resources. The *DFHG* provides a web API that can be queried with author names and fragment numbers. The result is a JSON output containing every piece of information about the requested fragment (e.g., <http://www.dfhg-project.org/DFHG/api.php?author=ACUSILAU&fragment=10>). The *DFHG* exports data to CSV and XML format files (both as EpiDoc XML and well formed XML).

- Integration with external resources.** One of the main goals of the project is to make the *DFHG* part of a bigger infrastructure of processed data. This is the reason why the *DFHG* is integrated with external resources such as textual collections, authority lists, dictionaries, lexica and gazetteers. These resources are fundamental for disambiguating and annotating *DFHG* data, which in turn offers a collection of parsed texts for enriching external libraries of Greek and Latin sources. The *DFHG* is currently connected to different resources that provide morpho-syntactic information and named entities disambiguation of textual data of the *FHG*. The *DFHG* provides also a *Müller-Jacoby Table of Concordance*, which is a complete correspondence between fragmentary historians published in the *FHG* and in *Die Fragmente der griechischen Historiker* including the *continuatio* and the *Brill's New Jacoby* (<http://www.dfhg-project.org/Mueller-Jacoby-Concordance>). The goal of this resource is to go beyond the *FHG* corpus and produce a complete catalog of fragmentary authors of Greek literature published in different digital editions. This resource is progressively ingested into the *Perseus Catalog* (<http://catalog.perseus.org>).
- Data citation.** It is possible to retrieve and export citations of *DFHG* fragments and source texts down to the word level using URN identifiers. These URNs are combinable with a URL prefix (<http://www.dfhg-project.org/DFHG/#>) to generate stable links. The syntax of each URN represents the editorial work of Karl Müller, who has arranged the fragments in a sequence and has attributed them to fragmentary authors, works, work sections and book numbers (e.g., urn:lofts:fhg.1.hecataeus.hecataei_fragmenta.genealogiae.liber_secundus:350). The *DFHG* provides also CITE URNs according to the guidelines of the CITE Architecture (<http://cite-architecture.org/>).
- Source Catalogs.** The *DFHG* includes a *Fragmentary Authors Catalog* and a *Witnesses Catalog* that have been created from *FHG* data. These catalogs allow users to search and visualize the 636 Greek fragmentary historians of the collection and each of their witnesses (i.e., authors who preserve quotations and text reuses of the fragmentary historians). Data from both catalogs has been used to generate charts for visualizing chronological distributions and statistics of *FHG* authors and their source texts. This data integrates also *Pleiades* identifiers with geo-locations that have been used for producing maps that visualize the geographical distribution of *FHG* authors and their witnesses.
- Text Reuse Detection.** The *DFHG* project offers experimental text reuse functionalities for automatic text reuse detection of *FHG* fragmentary historians. This resource allows users to automatically detect text reuses (fragmenta) of *FHG* authors in their witnesses. Users can insert an XML file URL or select one of the *PerseusDL* or *Open Greek and Latin* editions available in the *DFHG*. Results display quotations and text reuses of *FHG* authors within their source texts. The *DFHG* allows scholars to download complete XML files of the source texts of the fragments with dfhg attributes that mark up the presence of *DFHG* text reuses in the relevant passages of the source texts. *DFHG* text reuse detection is based on the Smith-Waterman algorithm that performs local sequence alignment to detect similarities between strings.
- OCR Editing.** The digital version of the *DFHG* has been produced starting from the OCR output of the printed edition of the *FHG*. Even if it is possible to obtain very good results when OCRing 19th-century editions of ancient Greek and Latin sources, OCRed texts still contain errors. The *DFHG* offers an interface for manual OCR correction of source texts, fragments, Latin translations and commentaries. Corrections are validated or rejected by the project team through an administration page.

Further developments of the *DFHG* project aim at implementing named entities recognition in the texts of Greek and Latin *fragmenta* and in contributing to enrich the number of lemmata and inflected forms of Greek and Latin thesauri. The final goal of the project is to offer a new methodology based on digital and computational approaches to represent complex historical text reuse data. The *DFHG* also offers an open collection of quotations and text reuses of Greek fragmentary historians. This resource provides the community of scholars and students with machine processable data for historical and computational research.

Appendix A

Bibliography

1. **Berti, M.** (2018). “Annotating Text Reuse within the Context: the *Leipzig Open Fragmentary Texts Series* (LOFTS)”. In Tischer, U., Gärtner, U. and Forst, A. (eds), *Text, Kontext, Kontextualisierung. Moderne Kontextkonzepte und antike Literatur*. Hildesheim, Zürich, and New York: Olms, 223-234.
2. **Berti, M.** (2019). “Historical Fragmentary Texts in the Digital Age”. In Berti, M. (ed), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*. Berlin and Boston: De Gruyter, 257-276. doi: 10.1515/9783110599572-015
3. **Berti, M., Almas, B. and Crane, G.R.** (2016). “The *Leipzig Open Fragmentary Texts Series* (LOFTS)”. In Bernstein, N.W. and Coffee, N. (eds), *Digital Methods and Classical Studies*. DHQ Themed Issue 10(2). <http://www.digitalhumanities.org/dhq/vol/10/2/000245/000245.html> (accessed 13 April 2019)
4. **Berti, M., Almas, B., Dubin, D., Franzini, G., Stoyanova, S. and Crane, G.R.** (2014-2015). “The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors”. *Journal of the Text Encoding Initiative* 8. doi: 10.4000/jtei.1218
5. **Berti, M., Blackwell, C. W., Daniels, M., Strickland, S. and Vincent-Dobbins, K.** (2016). “Documenting Homeric Text-Reuse in the *Deipnosophistae* of Athenaeus of Naucratis”. In Bodard, G., Broux, Y. and Tarte, S. (eds), *Digital Approaches and the Ancient World*. BICS Themed Issue 59(2): 121-139. doi: 10.1111/j.2041-5370.2016.12042.x
6. **Berti, M., Crane, G. R., Yousef, T., Bizzoni, Y., Boschetti, F. and Del Gratta, R.** (2016). “*Ancient Greek WordNet* Meets the *Dynamic Lexicon*: the Example of the Fragments of the Greek Historians”. In Mititelu, V.B., Forăscu, C., Fellbaum, C. and Vossen, P. (eds), *Proceedings of the Eighth Global WordNet Conference, Bucharest, Romania, January 27-30*. Bucharest, 34-38.

Monica Berti (monica.berti@uni-leipzig.de), University of Leipzig, Germany