Towards the Automatic Retrieval of Cited Parallel Passages from Secondary Literature

Matteo Romanello
(DAI/EPFL)

Publications in Classics constitute an ideal use case for information extraction technologies as they contain a high density of references to a variety of sources: references to canonical and fragmentary texts, inscriptions, manuscripts, papyri, museum objects, coins etc. The distinctive characteristic of canonical references, as opposed to other kinds of bibliographic references, is that they refer to an abstract division of the text into more granular units (e.g. chapters, verses, etc.). In Classics, as also in other disciplines, such references constitute the standard way used by scholars to refer to ancient (canonical) texts. Services for dealing with this kind of references were deemed to be one of the necessary components of a cyberinfrastructure for Classics.[1] Over the last decade, many efforts were put into building a digital infrastructure capable of dealing with canonical texts and references. At its core there is the Canonical Texts Services protocol (CTS), a network protocol for "translating" references into machine-actionable links.[2]

In this talk I focus specifically on the automatic extraction of canonical references from publications, which builds upon the other components of this infrastructure, and most importanly the CTS protocol.[3] In particular, I discuss two applications that are enabled by the availability of these citation data: firstly, the study of scholarly reception in Classics and, secondly, the retrieval of cited parallel passages from secondary literature. The assumption behind the use of automatically extracted canonical references to study scholarly reception is that the citation of a passage tells us, at the very least, that the cited text was used by some scholar at a certain point in time. Based on this assumption the diachronic frequency of such references can be used to track – to some extent – the varying scholarly fortune of classical authors over time.

The second application is more closely related to the information retrieval needs of classicists. Since parallel passages – loci paralleli – are signalled within publications by means of canonical references, extracting the latter means also enabling readers to search for the former. By using examples drawn from JSTOR, from which I extracted automatically all canonical references, I show how it becomes possible for the reader to search e.g. for all articles that relate, within the same sentence (or sentence window), a Vergilian passage to a Homeric passage. Such a functionality could be particularly useful to anyone researching topics related to intertextuality as well as to scholars preparing a commentary on a specific text.

References

Crane, Gregory, Brent Seales, and Melissa Terras. 2009. "Cyberinfrastructure for Classical Philology." *Digital Humanities Quarterly* 3 (1). http://www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html.

Romanello, Matteo. 2015. "From Index Locorum to Citation Network: an Approach to the Automatic Extraction of Canonical References and Its Applications to the Study of Classical Texts." Ph.D. thesis, King's College London. doi:11858/00-1780-0000-002A-4537-A. http://hdl.handle.net/11858/00-1780-0000-002A-4537-A.

[1] Crane, Seales, and Terras (2009).
[2] See e.g. Smith and Blackwell (2012).
[3] See Romanello (2015).

Smith, Neel, and C. W. Blackwell. 2012. "Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture." In *Donum Natalicium Digitaliter Confectum Gregorio Nagy Septuagenario a Discipulis Collegis Familiaribus Oblatum: A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends*, edited by Victor Bers, David Elmer, Douglas Frame, and Leonard Muellner. Washington, D.C.: Center for Hellenic Studies.
http://chs.harvard.edu/CHS/article/display/4846.