Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead

Giovanni Colavizza and Matteo Romanello

Journal of European Periodical Studies, 4.1 (Summer 2019)

ISSN 2506-6587

The *Journal of European Periodical Studies* is hosted by Ghent University

Website: ojs.ugent.be/jeps

To cite this article: Giovanni Colavizza and Matteo Romanello, 'Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead', *Journal of European Periodical Studies*, 4.1 (Summer 2019), 36–53

Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead

Giovanni Colavizza and Matteo Romanello
University of Amsterdam and École Polytechnique Fédérale de Lausanne
g.colavizza@uva.nl and matteo.romanello@epfl.ch

ABSTRACT

Even large citation indexes such as the Web of Science, Scopus, or Google Scholar cover only a small fraction of the literature in the humanities. This coverage sensibly decreases going backwards in time. Citation mining of humanities publications — defined as an instance of bibliometric data mining and as a means to the end of building comprehensive citation indexes — remains an open problem. In this contribution, we discuss the results of two recent projects in this area: *Cited Loci* and *Linked Books*. The former focused on the domain of classics, using journal articles in JSTOR as a corpus; the latter considered the historiography on Venice and a novel corpus of journals and monographs. Both projects attempted to mine citations of all kinds — abbreviated and not, to all types of sources, including primary sources — and considered a wide time span — nineteenth to twenty-first century. We first discuss the current state of research in citation mining of humanities publications. We then present the various steps involved in this process, from corpus selection to data publication, discussing the peculiarities of the humanities. The approaches taken by the two projects are compared, allowing us to highlight disciplinary differences and commonalities, as well as shared challenges between historiography and classics in this respect. The resulting picture portrays humanities citation mining as an area of research with a great, yet mostly untapped, potential and a few open challenges. Its potential lies in using citations to interconnect digitized collections on a large scale, by making explicit the *linking function* of bibliographic citations. As for the open challenges, a key issue is the existing need for an integrated metadata infrastructure, and an appropriate legal copyrights framework to facilitate citation mining in the humanities.

KEYWORDS

Bibliometrics, citations, citation indexing, information retrieval, Classics, History

Introduction

Scientific research is a joint and cumulative effort, where each contribution made needs to be informed by, and embedded into, previous literature. Yet such is the amount of scientific literature, that it would be impractical for scholars to work without the aid of citation indexes: search engines, such as Google Scholar, that allow for navigating the network of relations that the scientific literature forms. Citation mining is a core component of any citation index, and can be defined as the task of extracting relations between citing publications and the sources they cite. A citation is therefore a relation between two identified sources. Usually, citations are expressed in the text of the citing publication as references, often located in footnotes and/or reference lists. The references, or the short portions of a text where a source is cited, follow specific format rules which aim at unambiguously identifying the cited source, in the least possible amount of text. A citing source can contain multiple references to the same cited source, which we call in-text references or mentions. To be sure, the importance of citation indexes, and consequently of citation mining, goes beyond information retrieval. Citations are also used, among other things, for understanding science from a sociological and bibliometric perspective, and to track the acknowledgement of scientific credit for the purpose of evaluation.

Citation mining is by now a 'solved problem' with respect to STEM literature. Despite some drawbacks and margins for improvement, the crawlers behind mainstream indexes such as Google Scholar, Web of Science, Scopus, and Dimensions are largely capable of mining most citations accurately.[1] The main problem which is left open is the skewness in literature coverage and mining performance over different disciplines, with those within the humanities usually faring worse than most.[2]

Several reasons for this have been identified, which can be grouped into two categories:

- *Intrinsic factors*, which depend on the characteristics of the literature published in the humanities: lower proportion of born digital or digitized publications, higher variety of languages and publication venues, of publication typologies (monographs, articles, contributions, etc.), of referencing practices (morphology, syntax, and semantics of citations, including references usually made in footnotes, not just reference lists), the richness of referencing motivations.[3]
- *Extrinsic factors*, which depend on the information environment where citation mining is performed, mainly including the variety and fragmentation of supporting catalogues and information systems for unique identifiers and authoritative metadata.

1 Philippe Mongeon and Adèle Paul-Hus, 'The Journal Coverage of Web of Science and Scopus: A Comparative Analysis', *Scientometrics*, 106.1 (2016), 213–28; Gali Halevi, Henk Moed, and Judit Bar-Ilan, 'Suitability of Google Scholar as a Source of Scientific Information and as a Source of Data for Scientific Evaluation: Review of the Literature', *Journal of Informetrics*, 11.3 (2017), 823–34; and Daniel W. Hook, Simon J. Porter, and Christian Herzog, 'Dimensions: Building Context for Search and Evaluation', *Frontiers in Research Metrics and Analytics,* vol. 3 (2018).

2 Anne-Wil Harzing and Satu Alakangas, 'Google Scholar, Scopus, and the Web of Science: A Longitudinal and Cross-Disciplinary Comparison', *Scientometrics*, 106.2 (2016), 787–804.

3 Anton J. Nederhof, 'Bibliometric Monitoring of Research Performance in the Social Sciences and the Humanities: A Review', *Scientometrics*, 66.1 (2006), 81–100; Mu-hsuan Huang and Yu-wei Chang, 'Characteristics of Research Output in Social Sciences and Humanities: From a Research Evaluation Perspective', *Journal of the American Society for Information Science and Technology*, 59.11 (2008), 1819–28; and Chris A. Sula and Matthew Miller, 'Citations, Contexts, and Humanistic Discourse: Toward Automatic Extraction and Classification', *Literary and Linguistic Computing*, 29.3 (2014), 452–64.

We compare here two recent projects — *Cited Loci* and *Linked Books* — which attempted citation mining in two different areas of the humanities: classics and history, respectively. Despite their differences, both projects highlight the feasibility of the task and its subsequent importance for the field. We take the opportunity to discuss a typical citation mining pipeline, and present the specific challenges in the humanities. *Cited Loci* and *Linked Books* showcase how the availability of citation data in the humanities can greatly contribute to these disciplines in a variety of ways: information retrieval, understanding of knowledge accumulation processes and disciplinary historical dynamics, understanding the relative importance of different sources,[4] and studying the scholarly reception of literary authors.[5]

The comparison of the two projects discussed here further allows us to highlight a few key areas for improvement, which we deem critical to solving the broader problem of citation mining of scholarly literature in the humanities: digitization, availability of data (metadata, images, text), legal copyright frameworks, standards for representing and exchanging data (metadata, images, text and citations).

This contribution is organized as follows: we briefly discuss the state of the art regarding citation mining and indexes for the humanities. We then present the two projects, *Cited Loci* and *Linked Books*, and sketch a framework for their comparison, which mimics the choices we faced during both projects. We then compare the projects using the framework, and conclude by discussing what we consider the key areas of development in this domain.

State of the Art

The coverage of mainstream citation indexes is broad and constantly improving over time,[6] both for journals[7] and monographs,[8] despite some known limitations. The main problem which is left open, with respect to our focus here, is the skewness in literature coverage and mining performance over different disciplines, with humanities disciplines usually faring worse than most.[9] The resulting lack of citation data in the humanities remains a known problem, lamented several times over.[10] For these and other reasons, the use of citations to evaluate research in the humanities has also been questioned:[11]

4 Giovanni Colavizza, 'The Core Literature of the Historians of Venice', *Frontiers in Digital Humanities*, 4.14 (2017); Giovanni Colavizza, Matteo Romanello, and Frédéric Kaplan, 'The References of References: A Method to Enrich Humanities Library Catalogs with Citation Data', *International Journal on Digital Libraries*, 19.2–3 (2018), 151–61.

5 Matteo Romanello, 'Large-Scale Extraction of Canonical References: Challenges and Prospects', preprint (2018).

6 John Mingers and Loet Leydesdorff, 'A Review of Theory and Practice in Scientometrics', *European Journal of Operational Research*, 246.1 (2015), 1–19; Ludo Waltman, 'A Review of the Literature on Citation Impact Indicators', *Journal of Informetrics*, 10.2 (2016), 365–91; and Halevi and Bar-Ilan.

7 Mongeon and Paul-Hus.

8 Alesia Zuccala and others, 'Can we Rank Scholarly Book Publishers? A Bibliometric Experiment with the Field of History', *Journal of the Association for Information Science and Technology*, 66.7 (2015), 1333–47.

9 Harzing and Alakangas.

10 Richard Heinzkill, 'Characteristics of References in Selected Scholarly English Literary Journals', *The Library Quarterly*, 50.3 (1980), 352–65; A. J. M. Linmans, 'Why with Bibliometrics the Humanities Does not Need to Be the Weakest Link: Indicators for Research Evaluation Based on Citations, Library Holdings, and Productivity Measures', *Scientometrics*, 83.2 (2009), 337–54; and Sula and Miller.

11 Mike Thelwall and Maria M. Delgado, 'Arts and Humanities Research Evaluation: No Metrics Please just Data', *Journal of Documentation*, 71.4 (2015), 817–33; and Michael Ochsner, Sven E. Hug, and Hans-Dieter Daniel, 'Humanities Scholars' Conceptions of Research Quality', *Research Assessment in the Humanities*, ed. by Michael Ochsner, Sven E. Hug, and Hans-Dieter Daniel (Berlin: Springer, 2016), 43–69.

It appears clear that the availability of citation data would not completely solve the issue of research evaluation in the humanities.[12]

As a precondition of citation indexing, the automatic extraction of references from scholarly publications is a mature area of research. Recent developments include fully fledged architectures to extract and use citation data, embedded within digital library systems.[13] Several reference extraction services exist, such as ParsCit,[14] BILBO,[15] GROBID,[16] FreeCite, and CERMINE.[17] A recent survey and evaluation of several non-commercial reference parsing tools found that the best three performing ones all use Conditional Random Fields (CRF) as the supervised machine learning technique of choice: GROBID, CERMINE, and ParsCit, in order.[18] All three benefit from task-specific tuning using extra annotated data, with GROBID showing the best off-the-shelf results. Indeed, seven out of the thirteen surveyed tools use a CRF approach, while the rest mainly adopt regular expressions. The most recent literature on the topic employs CRF, Markov logic networks or deep learning.[19]

In summary, since referencing in the humanities is a less standardized practice than in other disciplines, there are consequences for the automatic extraction of citations. More specifically, reference lists at the end of a publication are not always given, as citations are commonly made in footnotes. Furthermore, humanists developed elaborate practices for the abbreviation and encoding of references, which also entail using a variety of formatting features such as italics or variations in type module. Eventually, it is common in the humanities to refer to both primary and secondary sources. The variety of cited materials and their physical existence in a multiplicity of collections results in a still fragmentary information ecosystem, with respect to their metadata. Unfortunately, these characteristics of the literature and sources in the humanities make it difficult to reuse existing services out-of-the-box.

12 Björn Hammarfelt, 'Four Claims on Research Assessment and Metric Use in the Humanities', *Bulletin of the Association for Information Science and Technology*, 43.5 (2017), 33–8.

13 Jlian Wu and others, 'Citeseerx: AI in a Digital Library Search Engine', *Innovative Applications of AI Conference* (2017), 2930–37.

14 Isaac G. Councill, Lee C. Giles, and Min-Yen Kan, 'ParsCit: An Open-Source CRF Reference String Parsing Package', *Proceedings of the Language Resources and Evaluation Conference* (2008), pp. 661–67.

15 Young-Min Kim and others, 'Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles, and Blogs', *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media, and Crowdsourcing* (2011), pp. 41–48.

16 Patrice Lopez, 'GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications', *Research and Advanced Technology for Digital Libraries*, (Berlin, Heidelberg: Springer, 2009), 473–74.

17 Dominika Tkaczyk and others, 'CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature', *International Journal on Document Analysis and Recognition*, 18.4 (2015), 317–35.

18 Dominika Tkaczyk and others, 'Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers', *Proceedings of ACM JCDL* (2018), pp. 1–10.

19 See respectively Martin Körner and others, 'Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications', in *New Trends in Databases and Information Systems* (Cham: Springer, 2017), pp. 137–45; Dustin Heckmann and others, 'Citation Segmentation from Sparse and Noisy Data: A Joint Inference Approach with Markov Logic Networks', *Digital Scholarship in the Humanities*, 31.2 (2016), 333–56; and Danny Alves Rodrigues, Giovanni Colavizza, and Frédéric Kaplan, 'Deep Reference Mining from Scholarly Literature in the Arts and Humanities', *Frontiers in Research Metrics and Analytics*, vol. 3 (2018).

Cited Loci and Linked Books

The project Cited Loci[20] (CL hereafter) aims at extracting references to classical authors — the so-called *canonical references* — from articles in JSTOR.[21] Alongside references to papyri, inscriptions, manuscripts, and museum objects, these references play a key role as they point to the very object of study, namely classical texts. What makes canonical references special — and worthy of investigation by computational methods — is that they have been essentially stable for the last three centuries. Based on these considerations, the project has a two-fold goal: First, to develop better means for information retrieval in this field by leveraging the automatically extracted citation data; and, second, to study the fortune of classical authors by using canonical references as a proxy of the attention that classical authors received by scholars over time.

The project Linked Books (LB hereafter) aims at creating a comprehensive citation index of the historiography on Venice, considering all cited sources, including primary sources such as documents held at the Archive of Venice.[22] This project, unable to rely on pre-existing digital resources, started with a digitization campaign to acquire its corpus. LB was organized into two tracks: Data and products, and bibliometrics. The first part was devoted to the release of citation mining methods and citation data, and to the development of two interfaces: A digital library and a citation index. The second track was devoted to the bibliometric study of the historiography of Venice, and historiography as a discipline more broadly.

The Citation Mining Framework

To be able to compare the two projects, we defined a more general framework for citation mining, encompassing the various steps of the process as illustrated in Fig. 1.

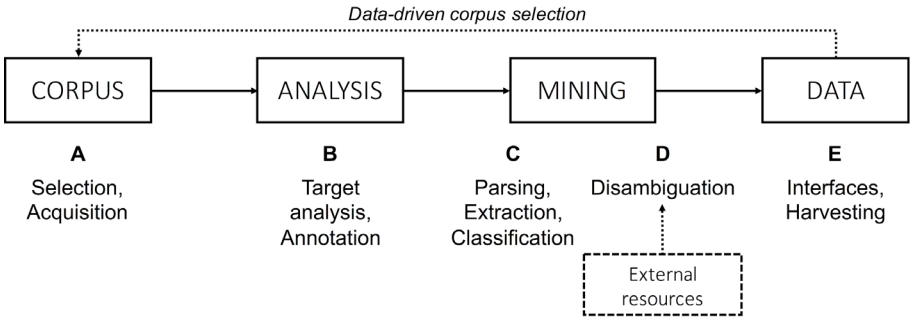

Fig. 1 The citation mining framework used for comparison.

20 JSTOR collaborated in this project as a data provider, making available under the Data for Research scheme its dataset of classics-related journal articles.

21 See Matteo Romanello, *From Index Locorum to Citation Network: An Approach to the Automatic Extraction of Canonical References and its Applications to the Study of Classical Texts* (doctoral dissertation, King's College London, 2015).

22 The Linked Books project is a collaboration among several partner institutions: the École Polytechnique Fédérale de Lausanne (EPFL); the Ca' Foscari University of Venice, especially its Humanities Library (BAUM); the National Library of St. Mark; the State Archive of Venice; the Venetian Institute for Sciences, Letters and Arts; the European Library of Information and Culture (BEIC); and the Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU).

Corpus selection and acquisition

The first step is the most decisive one, as choices made at this stage concerning the composition of the corpus will have waterfall effects on all subsequent steps. In an ideal situation, a corpus already exists that can be mined for citations. Unfortunately, often this is not the case, especially with humanities publications. If the corpus needs to be created from scratch, strategies may vary considerably depending on the subject. Some research areas, such as the long-established ones, are easier to define in terms of their publications (i.e. it is relatively easy to identify books and serial publications belonging to such areas). In other cases, such as the history of Venice, it may be more sensible to leave the corpus open, and define it in an incremental and iterative fashion.

If a new corpus needs to be created, there are several aspects that need to be carefully considered. Firstly, the state of digitization of the materials involved: If they have not been already digitized, or only to a small extent, it is necessary to set up a pipeline to digitize the materials, to assign descriptive metadata as required, and to perform optical character recognition (OCR) on the images. Secondly, it is important to verify the provision of metadata and their formats: Libraries are most likely to already have some metadata, but they may be in a format not yet readily usable in a citation mining project. Or, metadata formats may be very heterogeneous, especially when gathered from several libraries and different countries, thus requiring some mapping to a common format. Last but not least important, is the legal context around the digitized materials. It is essential to clarify from the outset with the data providers who will have what rights to the digitized materials, and which rights will apply to derived materials (e.g. annotations, be they manual or automatic).

Target analysis and annotation

Once the corpus has been selected, acquired, or created, the next step is to carry out a preliminary campaign aimed at manually annotating the pieces of information which are to be automatically identified. The annotation activity allows to assess where bibliographic references are most often found — references lists, footnotes — and whether their stylistic variations follow specific patterns — e.g. by author, publisher, publication venue, or period. This step provides an understanding of the materials, of their coherence and variations, and it is especially important if the publications in the corpus cover a wide diachronic span.

The creation of a golden set (or ground truth) — a set of carefully checked annotations — is necessary in order to be able to evaluate the accuracy of the citation extraction results and, if a supervised machine learning approach to the extraction is taken, a substantial amount of such annotations is necessary to train the algorithm of choice.

Parsing, extraction, classification

The task of mining citations from texts consists of two subsequent steps, namely citation extraction and citation matching (or disambiguation). Citation extraction, similarly to Named Entity Recognition, is a sequence labelling — or structured prediction — problem, as it requires to label tokens in a text as part of a bibliographic reference (or not). Yet, the very same problem can be approached in a variety of ways and the viability and suitability of each approach — or the need for ad-hoc solutions — will clearly appear after the target analysis and manual annotation steps. If little or no annotated data are available, it is advisable to adopt an existing system or to rely on a rule-based

approach. Conversely, if there is no shortage of annotated data, it is possible to go for a fully supervised approach as well as to consider hybrid systems, which use a combination of manually defined rules and trained models.

Disambiguation and resources

Once bibliographic references have been extracted, citation matching seeks to establish whether any two references are pointing to the same publication. This allows us to move from references to citations, a necessary step to derive any quantitative indicators from citation data.

With respect to citation extraction, the problem of matching references can be tackled with a variety of approaches, ranging from direct or rule-based lookup, to supervised and unsupervised machine learning. Bibliographic metadata also play a key role in this context: in order to disambiguate citations, it is essential to have unique identifiers for citable objects that can be assigned to citations, for example DOIs. Library catalogues, as well as institutional information systems, are valuable sources for such identifiers, yet it is important to verify the accessibility of their metadata, e.g. whether Application Programming Interfaces (APIs) are provided.

Publication and (re-)use of project outputs

As a result, this framework generates citation data and reference annotations — as a by-product. Interactive tools are required in order to make this data explorable and searchable by users, while interfaces are needed to expose this data and make it reusable by others for potentially different purposes. The legal conditions under which the publication corpus is made available determine to a large extent which portions of the full text can be released as part of the data.

A Side-by-Side Comparison of the two Projects

This section compares the two projects following the proposed general framework describing a simplified pipeline to build a citation index: Corpus selection and acquisition, target analysis, parsing, extraction and classification of references, disambiguation of citations, publication and reuse of citation data.

Corpus Selection and Acquisition

Cited Loci

Due to internal time and resource constraints, CL's corpus coincides with the journal articles contained in JSTOR and automatically classified as belonging to the field of Classical Studies. The corpus consists of a snapshot of JSTOR's content, taken in 2013 and including 138,821 articles in plain text format. Three million references to ancient authors, works, and specific text passages were automatically extracted.

As noted above, the CL project used as a corpus the classics articles in JSTOR, whose contents are made freely accessible for research purposes under the Data for Research program. High quality article-level metadata — an essential element for the analysis of the extracted citation data — were obtained via JSTOR's APIs. The full-text articles were provided, first as plain text files and, subsequently, as OCR with image coordinates encoded in JSON.

The choice of JSTOR as a corpus was a pragmatic one, the main advantage being that JSTOR is 'ready-to-mine'. No additional work was needed to gather article-level metadata or scrape the full-text from web pages or online PDFs, as is the case, for example, if one decides to use the journals listed in the Ancient World Online (AWOL) index.

However, there are disadvantages to this choice, most notably:

- One has no direct influence over the OCR process and quality of its result. This would be desirable especially for ancient Greek texts contained in the articles, given that the OCR quality varies greatly depending, for example, on age of publication and language.
- One also has no control over the construction of the corpus, especially its representativeness, although its coverage can be determined and quantified.
- Finally, the JSTOR data are not openly available, meaning that derivative materials — such as annotated data for training/evaluation — can be openly published and shared only insofar as they do not repurpose a substantial amount of an article's text, and given approval of JSTOR's legal office.

Linked Books

The historiography on a topic, such as the history of Venice, is not an easily bounded area of research. A variety of contributions exist, with no existing database even close to being representative of the field. Consequently, LB's corpus has been considered as unbounded from the very beginning, and to be progressively enlarged using mined citation data to regularly assess its coverage. A first selection of literature was individuated from scratch by using a combination of: library catalog lookup (search by topic and keyword), domain expert advice, existing published bibliographies, rapid-access shelves in specialized research libraries. This selection comprises circa 2,000 monographs and 10 local journals mostly in Italian (currently 552 issues, for a total of 5,496 articles), and it was used as a seed for the first digitization campaign. Newly published or particularly well-cited literature can be digitized and indexed during subsequent campaigns.

Once a selection of the relevant literature is made, it can be digitized and its catalog metadata acquired. Subsequently, the images are OCRed, in our case using a commercial solution tuned for the specific materials at hand. Given that LB operates from scratch, specific decisions could be taken for the following interrelated areas:

1. *Copyright*: we established partnership with libraries in possession of the literature, which was made available under the agreement that a digital copy was to be given back to the holding library, that only temporary copies could be used for the purpose of reference mining, and that references and citations, once extracted, would not per se constitute a violation of the publisher's or author's copyright, not constituting an integral part of the contents.
2. *Metadata*: they were acquired from the Italian National Catalog and corrected or complemented as necessary. For example, an important piece of information is the library provenance of an item, for copyright and source verification purposes.
3. *Digitization and OCR*: we could make choices in view of subsequent needs. For example, during OCR we also extracted layout features (such as font size and usage of italics or bold, often used in footnote references), in order to use this information for citation mining.

The resulting output was a collection of paired JPG and HTML files, one for each page of an item.[23]

Target Analysis and Annotation

Cited Loci

The wide variety of sources that are cited within classics publications makes them a perfect testbed to develop and test information extraction systems. In addition to bibliographic references pointing to modern publications (i.e. secondary sources) they contain references to various kinds of primary source materials, including:

- canonical texts (e.g. "Virgil, *Aen.* 12, 101-109, Hom. *Il.* 7.180");
- fragmentary texts (e.g. "FGrHist 688 F 13; Alc. fr. 34,1 Voigt");
- inscriptions (e.g. "CIL 3, 6174; AE 1991, 1405");
- papyri (e.g. "PCair. inv. 10750");
- manuscripts (e.g. "Vendôme, Bibl. mun. 31");
- museum objects (e.g. coins, vases, etc.).

Of all these sources, the CL project focused exclusively on references to canonical texts, where *canonical* means that these texts can be referred to by means of a citation scheme agreed upon by scholars. This is the case with canonical texts, but not with fragmentary texts or inscriptions, where texts always need to be cited according to a specific (critical) edition.

The main reason for narrowing down the project's focus to canonical references, in addition to the already mentioned constraints of time and resources, is the limited availability of unique and persistent identifiers that are needed for the disambiguation of other reference types. In fact, a single source of unique, machine-readable identifiers for all extant papyri or inscriptions is still lacking, thus hindering further advancements on the automatic extraction of these references.[24] On the contrary, for canonical texts we can rely on two already established and comprehensive resources, namely the Perseus Catalog and the Classical World Knowledge Base. The two resources were combined into a knowledge base especially designed to support the extraction and disambiguation of canonical references, the HuCit Knowledge Base.

A *reference style* is a specific combination and encoding of elements in a reference. Variations in the reference styles can be attributed essentially to three factors:

1. *Language*: the abbreviations of ancient authors and works may slightly vary from one language to another.
2. *Date of publication*: some citation practices – such as the use of Roman numerals to indicate book numbers – seem to be more or less common, depending on the period when the article was published.
3. *Target audience*: articles written for a highly specialized audience tend to display more concise (and thus obscure) abbreviations in the references, whereas in

23 The approach taken for corpus selection and acquisition is thoroughly described with more details in Colavizza, Romanello, and Kaplan, 'The References of References'.

24 However, there exist various resources that provide stable uniform resource identifiers (URIs) for several kinds of materials that are referred to within Classics publications: for example, the Eagle Europeana portal for inscriptions, Papyri.info for papyri; the Arachne database for archaeological objects; and the Leipzig Open Fragmentary Texts Series (LOFTS) for fragmentary texts.

articles written for a more general audience it is customary to spell out the name of the author and the title of the work cited.

As for the annotation of canonical references, it was not feasible to carry out an ad-hoc annotation campaign for the Cited Loci project. Instead, the system components that are based on machine learning (and thus require a training set) relied on a corpus of bibliographic abstracts where canonical references were annotated. This dataset consists of abstracts drawn from *L'Année Philologique* (APh) — the reference bibliography for classical studies — and written in English, French, German, Italian, and Spanish.[25] However, the inherent differences between the APh abstracts and the articles in JSTOR would have justified the creation of dedicated training/test sets. In fact, texts greatly differ in terms of document length as well as citation styles that are represented (the APh corpus is very homogeneous, whereas JSTOR contains a much wider variety of styles).

Linked Books

The aim of LB was to index all cited sources, in all their forms. The main general typologies of cited sources, with respect to the structure of their references, are:

1. *Primary sources*: any documentary evidence, either in original or edited, or non-scholarly publications.
2. *Secondary sources, books*.
3. *Secondary sources, articles, and contributions*: any publication contained within another publication, such as edited volume, journal issue, and the like. In this case, references contain both the extremes of the specific publication and of its container publication.

   Furthermore, references can be given in full, or in a variety of abbreviated forms, highly dependent on context. For example, the full primary source reference:

   > "Archive of Venice, *Procuratori di San Marco*, *de citra*, *commissarie*, b. 1, c. 7."
   > (Components: archive, record group, series, sub-series, box, sheet.)
   >
   > Can be abbreviated as:
   > A: "ASVe, PSM, *de citra, commissarie*, b. 1, c. 7."
   >
   > With acronyms defined at the beginning of the publication, or even:
   > B: "Ivi, c. 8."

To refer to the very same box as in the immediately previous reference, but another sheet. The procedure is similar for every kind of cited source. We identify A as a *global abbreviation* (dependent on the global reference context of the publication) and B as a *local abbreviation* (dependent on the local reference context, i.e. previous references). Given the variety of the literature, it was profitable to conduct a *reference style classification* to inform the choice of techniques to adopt in subsequent steps. This classification step entails considering samples of references, representative of different publication venues, publishers and periods, in order to find broad categories of reference styles.

In LB, a preliminary and explorative annotation campaign was conducted, focusing on the broadest variety of publications possible. This campaign allowed to

25 APh Corpus version 2.0.

establish a classification of cited sources and their abbreviations (given above), and a classification taxonomy for the components of references (author, title, year, archive, etc.). This campaign went on until a) major changes to the taxonomy no longer occurred and b) every element in the taxonomy was reasonably represented in terms of number of occurrences. Note that an estimate of the relative importance of each element in the taxonomy is relevant for its consolidation during parsing. Afterwards, all these preliminary annotations were discarded.[26] Lastly, this phase ended with an extensive annotation campaign, on a sample of the corpus, which yielded a golden set of over 40,000 manual annotations to be used for parsing.[27]

Parsing, Extraction, Classification

Cited Loci

The extraction and disambiguation of canonical references were modelled as a three-step process consisting of the following:

1. *Extraction of named entities*: a) names of ancient authors (e.g. Vergil); b) titles of works (e.g. *Aeneid*) and c) references to specific text passages (e.g. "Verg., *Aen.* 1,33").
2. *Detection of relations between entities*: since a reference is represented as a relation between two entities, the canonical references are reconstructed from the entities found in the text. For example, the reference "Verg. *Aen.* 1, 33" is expressed as a relation between the entity identifying the cited text (in this case "Verg. *Aen.*") and the entity indicating the citation scope ("1, 33"), namely the precise text passage being cited.
3. *Disambiguation of named entities and relations*: determining which authors, works and passages are referred to in the text is done by assigning a unique identifier to each entity and relation. The reference in the example above, for instance, will be assigned the uniform resource name (URN) "urn:cts:latinLit:phi0690.phi003:1.33". This identifier is built by concatenating the URN for the cited work (urn:cts:latinLit:phi0690.phi003 for Virgil's *Aeneid*) with a value representing the cited passage (1.33 which stands for book 1, line 33).

Representing references as relations between entities (as opposed to 'monolithic' named entities) makes it possible to handle consecutive references, as well as discursive references where the components of a reference may be found far from each other. Consider the following passage (named entities are highlighted):

> The picture of Achilles and of the ***Iliad*** that emerges from the twenty explicit references in the first half of the ***Aeneid*** is almost totally negative. Achilles is the unyielding (inmitis, **1.30, 3.87**), ferocious (saevus, **1.458, 2.29**) warrior of ***Iliad*** **20 and 21** [...]

Here, various passages of the *Iliad* (*Il.* 1,30; 3.87; 1.458; 2.29), related to Achilles' portrait in the Homeric poem, are cited in a rather discursive fashion. Treating these references

26 The details of the resulting taxonomy are given in Colavizza, Romanello, and Kaplan and in Colavizza, 'The References of References', and in Colavizza and Romanello, 'Annotated References in the Historiography on Venice: 19th–21st Centuries', *Journal of Open Humanities Data*, 3.2 (2017), no page.

27 The annotation data set is published in Colavizza and Romanello, 'Annotated References'.

as *relations* (e.g. between "Iliad" and "1.30", etc.) gives us the flexibility needed to design an annotation scheme that can cope with such cases.

Linked Books

The structure of references cannot be understood without investing some time to manually extract and annotate a sample of references from the corpus. Reference annotation is, in effect, a key step in citation mining, as it allows us to:

- understand the structure of references;
- establish a classification taxonomy for reference parsing;
- prepare an annotated corpus for supervised learning.

Given the availability of annotated data, and having reached a stability of the annotation taxonomy, we used supervised learning methods for parsing, extraction and classification of references in LB, relying on an established method for these tasks: Conditional Random Fields. We framed the tasks as follows:

1. A first parser, considering the full text of every item in the collection, tagged individual tokens with *specific tags* (i.e. using the annotation taxonomy discussed above, and tags such as author, title, year of publication). The low amount of annotations for more rare tags required some consolidation of the taxonomy.
2. A second parser, also considering the output of the first one, tagged every token in the text as being outside, inside, beginning or ending a reference, plus assigning a *general typology* to it: primary source, secondary source (book), secondary source (article).

For example, the following footnote (number 5):

> (5) A.S.V., Provveditori sopra monasteri, b. 280; Riformatori dello Studio di Padova, f. 272.

is parsed at first yielding the following result:

> "(5)" *out of reference.*
> "A.S.V.," *archive.*
> "Provveditori sopra monasteri," *record group.*
> "b. 280;" *box.*
> "Riformatori dello Studio di Padova," *record group.*
> "f. 272." *sheet.*

Then it is parsed a second time yielding the following result:

> "(5)" *out of reference.*
> "A.S.V., Provveditori sopra monasteri, b. 280;" *primary source.*
> "Riformatori dello Studio di Padova, f. 272." *primary source.*

The end result is thus the extraction of two references, with their components and general category.

An explicit choice made at this level was to maximize the *recall* evaluation score, at the expense of *precision*. This approach yielded a high number of false positives, or

tokens tagged with specific tags and general typologies despite not being part of a reference. Examples are book or article titles or in-text mentions of authors and other named entities. Nevertheless, as we will discuss in what follows, using high recall at this step and high precision for the subsequent disambiguation task, results in a balanced pipeline at the end.[28]

Disambiguation and Resources

Cited Loci

As mentioned in the previous section, the disambiguation of canonical references is done by means of Uniform Resource Names (URNs) that follow the special syntax defined by the Canonical Texts Services (CTS) protocol. This protocol was developed in the framework of the Homer Multitext project and is used to translate canonical references into machine-actionable links.[29]

Working with canonical references, as opposed to working with e.g. archival documents, is easier on several respects: first, classical authors and works form a closed set (i.e. the outcome of a historical process of canonization) and, second, the way these texts are cited is to a large extent fixed and stable. Given these two characteristics, it is feasible (and reasonable) to gather as much information as possible about classical authors and their texts into a knowledge base, aimed at supporting the extraction and disambiguation of such references. At the time of writing, this knowledge base contains 3,400 name variants for over 1,500 unique authors and over 6,500 work variants for 5,200 unique works, in addition to CTS URNs for both authors and works.

There are currently two implementations of the disambiguation of canonical references. The first employs a fuzzy matching approach to match references against the knowledge base, while the second implementation uses a machine learning approach called Learning to Rank to select the most likely candidate from a list of disambiguation candidates retrieved from the knowledge base. In addition to being slightly more accurate than the former (+.5%), the latter implementation presents the main advantage of being more readily adaptable to other types of publications (e.g. books), provided that sufficient annotated data can be produced.

The main difficulty when disambiguating canonical references remains a phenomenon that we call *implicit topicalization*. This happens when the general topic of a publication constitutes an essential piece of information for the disambiguation of the references contained in this publication. Consider as an example a whole book written about Plato's *Republic.* At some point in the book the author will start omitting the author's name when citing passages of this dialogue (e.g. writing *Rep.* 426b instead of Plato, *Rep.* 426b), since this dialogue is the *implicit topic* of the publication. Classical commentaries offer another example of this phenomenon: in a commentary about a tragedy by Sophocles, for instance, references to other tragedies will appear in a form more concise than usual, as the subject of the commentary itself ensures the reader's ability to decipher these references. Capturing contextual information automatically remains a considerable technical challenge, especially for the disambiguation of such references, as it requires some surrogate of the context to be embedded into the model. In the machine learning-based implementation of our disambiguation algorithm, we

28 Precision is the number of true positives over the total number of positive results, either true or false. Recall is the number of true positives over the number of true positives plus false negatives. Further discussion of technical details can be found in Colavizza, Romanello, and Kaplan, 'The References of References'.

29 Neel Smith, 'Citation in Classical Studies', *Digital Humanities Quarterly*, 3.1 (2009).

try to overcome the effect of *implicit topicalization* by leveraging contextual cues like the mention of the cited author or work in the document title, and information about the entities that surround a given reference.

Linked Books

In the LB project we needed to disambiguate authors and sources (of the three typologies given above). We could rely on a set of resources to link references to their referred items. First of all, most citations to books can be matched using the Italian National Catalog,[30] which we acquired and deployed locally, thanks to our collaboration with ICCU. Yet, the problem remains open on how to regularly update our local instance with the live ICCU version. Secondly, we could rely on the information system of the Archive of Venice (SiASVe), which was likewise replicated locally, and where every record group and document series of this archive is named, indexed and described. Lastly, for authors we used the publicly available VIAF API. Nevertheless, several references to items not to be found in these two systems, such as other primary sources or journal articles, are left out. At the same time, searching the three systems is somewhat costly (the Italian Catalog alone contains, at the time of writing, slightly over 16 million records). For these reasons, the disambiguation task was approached as follows:

1. A first internal search is performed on already disambiguated items. If a match is found, the system stops.
2. If no match is found, a search is conducted on the three systems, if the general typology of the reference is appropriate. If a match is found, the system stops.
3. If no match is found, a new entry in the index is added.

Given this general approach, searches are performed with different methods for every system:

1. The internal lookup, the Italian Catalog lookup and the VIAF lookup use a combination of string and rule matching.
2. The SiASVe lookup uses a supervised multinomial logistic classifier, that rolls back to rule matching if the probability for a classification is low (and therefore the referred item was likely never mentioned as part of the training data). It is worth noting that the classifier relies on a large amount of manually corrected disambiguations.

Despite our efforts, the disambiguation for the LB project is still far from satisfying. The lack of either external authority systems, or the lack of APIs to access them, greatly hindered our efforts to rely on external resources. At the same time, the reliance on internal lookup is limited by the quality of reference data.[31]

Publication and (Re-)Use of Project Outputs

As citation data are scarce in the humanities, it is critical for both projects to share project outputs in various formats with the goal of fostering new research in this area.

30 <opac.sbn.it>.

31 More details on our approach to disambiguation and its results for what concerns secondary sources (books), can be found in Colavizza, Romanello, and Kaplan, 'The References of References'.

The legal context constitutes for both projects a constraint with regards to making publicly accessible the collections of digitized articles.

Cited Loci

The CL project has already made available under an open source license all software components, as well as necessary data, for mining canonical references from text. These consist of:

1. Two Python libraries for the extraction and parsing of canonical references: the Citation Extractor and the Citation Parser.
2. The APh corpus, a dataset consisting of APh abstracts that can be used for the development and, most importantly, the evaluation of other solutions for the extraction of canonical references.
3. The HuCit Knowledge Base which consists of data modelled in RDF as well as a programming interface to these data written in Python.

The publication of two additional project outputs is planned: first, a dataset containing all extracted references and, second, the prototype of a search interface allowing the reader to search through JSTOR by using references to classical authors as a search key criterion.

Linked Books

The LB project released two software applications, a set of datasets and all its data via an API:

1. The two interfaces are a digital library application (Scholar Library, not public), meant to be accessible from the internal network of partner libraries (who possess the original materials) and a citation index (Venice Scholar, freely available online). The two applications are intimately connected, and allow to browse items in the index, with their given and received citations, and contents in the catalog. The references which are the basis for the establishment of a citation relationship are all searchable, as are the contents of the digitized materials from which they were extracted (albeit only from partner libraries, in this latter case). In so doing, no black-box exists, and users can inspect any step in the citation mining process.
2. Some partial datasets have been published online, and especially the corpus of annotated references, along with the code used to train models for supervised reference parsing, extraction and classification.[32]
3. The citation data that can be explored through the Venice Scholar can also be accessed programmatically via the openly available Venice Scholar API.

Citation Mining for the Humanities: Open Challenges

Our experience with these two projects shows that the main open challenges for citation mining in the humanities relate to infrastructure as they concern: 1) the availability of digital corpora; 2) the accessibility of catalog metadata; 3) standardized formats to expose bibliographic metadata; 4) legal frameworks for the use and publication of data.

32 See Colavizza, 'The Core Literature of the Historians of Venice'.

Digital corpora that can be mined for citations are few, fragmented – much in the same way that the humanities are fragmented into very specialized sub-disciplines – and most often available behind paywalls. The situation is much better for international and recent publications, especially in English, but problems remain for older and national scholarly literatures.

Catalog metadata can be extremely useful in a citation mining pipeline, especially for the disambiguation of extracted references. In reality, however, they can hardly be exploited as catalogs and information systems are usually fragmented into national — and sometime even institutional — silos.

Metadata are still exposed in a way that lacks standardization and uniformity, and only occasionally they are exposed and disseminated by means of APIs. This situation results into perhaps the main bottleneck at the moment: library catalogs and archive information systems cannot be replicated without enormous resources, yet they remain, for the best part, hard to access.

Finally, copyright issues on digitized materials constitute a great obstacle to citation mining and, more generally, to text mining of scientific publications. In particular, legal frameworks limit the freedom to openly share citation data, as well as datasets to train and evaluate citation mining solutions. This issue extends not only to historical materials, but also to contemporary scholarly publications, and has led to the proposal of revising the laws on copyright (both in Europe and in the UK) to add an exception for text and data mining.[33]

Conclusions

We see a single, main direction for future work, with a focus on infrastructure rather than on research. In our view, there is a general, urgent need to harvest and expose publication data and metadata in a uniform and centralized way at a European level. This consideration applies to periodicals, as well as any other publication and source relevant to scholars in the humanities. Metadata and image data should be encoded following uniform standards, both at the local and national level, in order to be harvested and centrally exposed. This is, of course, the goal of projects such as Europeana, whose importance cannot be stressed enough for the many direct and indirect benefits that such an endeavour would generate. Given the availability of image and/or text data and metadata, the other challenges are surmountable with a collaborative and technical effort.

The need for such an infrastructure does not only apply to the historical holdings of libraries and archives, but also to the monographs and periodical publications that are constantly being published, both in printed and in electronic form. Such a mechanism for harvesting metadata at a fine level of granularity would allow third-party services to access, mine and index these publications, without the need for manual intervention. On this respect, currently ongoing infrastructure projects[34] have the chance of realizing such an infrastructure at a large-scale, with a potentially huge impact on citation mining for the humanities.

Giovanni Colavizza is Assistant Professor of Digital Humanities at the University of Amsterdam. He was previously part of the research engineering group at the Alan

33 An ongoing project on this very topic is 'The Future of Text and Data Mining'.

34 See Open Access in the European Research Area through Scholarly Communication (OPERAS) and High Integration of Research Monographs in the European Open Science Infrastructure (HIRMEOS).

Giovanni Colavizza is Assistant Professor of Digital Humanities at the University of Amsterdam. He was previously part of the research engineering group at the Alan Turing Institute and of the quantitative science studies group at the Centre for Science and Technology Studies (CWTS), Leiden University. He works on machine learning and data science applied to GLAM collections (Galleries, Archives, Libraries, and Museums), and on the use of computational methods in the humanities.

Matteo Romanello is a digital humanities specialist with particular experience and expertise in the areas of classics, archaeology, and history. He received his PhD from King's College London in 2015 with a thesis entitled From Index Locorum to Citation Network: An Approach to the Automatic Extraction of Canonical References and its Applications to the Study of Classical Texts. He is currently Research Scientist in the Digital Humanities Laboratory at the Ecole Polytechnique Fédérale de Lausanne (EPFL). His research interests include natural language processing and information extraction, especially their domain-specific applications; citation mining and analysis; and applications of semantic web technologies in the humanities

Bibliography

Councill, Isaac G., Lee C. Giles, and Min-Yen Kan, 'ParsCit: An Open-Source CRF Reference String Parsing Package', *Proceedings of the Language Resources and Evaluation Conference* (2008), pp. 661–67

Colavizza, Giovanni, 'The Core Literature of the Historians of Venice', *Frontiers in Digital Humanities*, 4.14 (2017), no page

——, and Matteo Romanello, 'Annotated References in the Historiography on Venice: 19th–21st Centuries', *Journal of Open Humanities Data*, 3.2 (2017), no page

——, Matteo Romanello, and Frédéric Kaplan, 'The References of References: A Method to Enrich Humanities Library Catalogs with Citation Data', *International Journal on Digital Libraries*, 19.2–3 (2018), 151–61

Halevi, Gali, Henk Moed, and Judit Bar-Ilan, 'Suitability of Google Scholar as a Source of Scientific Information and as a Source of Data for Scientific Evaluation: Review of the Literature', *Journal of Informetrics*, 11.3 (2017), 823–34

Hammarfelt, Björn, 'Four Claims on Research Assessment and Metric Use in the Humanities', *Bulletin of the Association for Information Science and Technology*, 43.5 (2017), 33–8

Harzing, Anne-Wil, and Satu Alakangas, 'Google Scholar, Scopus and the Web of Science: A Longitudinal and Cross-Disciplinary Comparison', *Scientometrics*, 106.2 (2016), 787–804

Heckmann, Dustin, and others, 'Citation Segmentation from Sparse and Noisy Data: A Joint Inference Approach with Markov Logic Networks', *Digital Scholarship in the Humanities*, 31.2 (2016), 333–56

Heinzkill, Richard, 'Characteristics of References in Selected Scholarly English Literary Journals', *The Library Quarterly*, 50.3 (1980), 352–65

Huang, Mu-hsuan, and Yu-wei Chang, 'Characteristics of Research Output in Social Sciences and Humanities: From a Research Evaluation Perspective', *Journal of the American Society for Information Science and Technology*, 59.11 (2008), 1819–28

Hook, Daniel W., Simon J. Porter, and Christian Herzog, 'Dimensions: Building Context for Search and Evaluation', *Frontiers in Research Metrics and Analytics*, vol. 3 (2018)

Kim, Young-Min, and others, 'Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles, and Blogs', *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media, and Crowdsourcing* (2011), pp. 41–48

Körner, Martin, and others, 'Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications', in *New Trends in Databases and Information Systems* (Cham: Springer, 2017), pp. 137–45

Linmans, A. J. M, 'Why with Bibliometrics the Humanities Does not Need to Be the Weakest Link: Indicators for Research Evaluation Based on Citations, Library Holdings, and Productivity Measures', *Scientometrics*, 83.2 (2009), 337–54

Lopez, Patrice, 'GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications', *Research and Advanced Technology for Digital Libraries*, (Berlin, Heidelberg: Springer, 2009), 473–74

Mingers, John, and Loet Leydesdorff, 'A Review of Theory and Practice in Scientometrics', *European Journal of Operational Research*, 246.1 (2015), 1–19

Mongeon, Philippe, and Adèle Paul-Hus, 'The Journal Coverage of Web of Science and Scopus: A Comparative Analysis', *Scientometrics,* 106.1 (2016), 213–28

Nederhof, Anton J., 'Bibliometric Monitoring of Research Performance in the Social Sciences and the Humanities: A Review', *Scientometrics*, 66.1 (2006), 81–100

Ochsner, Michael, Sven E. Hug, and Hans-Dieter Daniel, 'Humanities Scholars' Conceptions of Research Quality', in *Research Assessment in the Humanities*, ed. by Michael Ochsner, Sven E. Hug, and Hans-Dieter Daniel (Berlin: Springer, 2016), pp. 43–69

Rodrigues Danny Alves, Giovanni Colavizza, and Frédéric Kaplan, 'Deep Reference Mining from Scholarly Literature in the Arts and Humanities', *Frontiers in Research Metrics and Analytics*, vol. 3 (2018)

Romanello, Matteo, *From Index Locorum to Citation Network: An Approach to the Automatic Extraction of Canonical References and its Applications to the Study of Classical Texts* (doctoral thesis, King's College London, 2015)

——, 'Large-Scale Extraction of Canonical References: Challenges and Prospects', pre-print (2018), no page

Smith, Neel, 'Citation in Classical Studies', *Digital Humanities Quarterly*, 3.1 (2009).

Sula, Chris A., and Matthew Miller, 'Citations, Contexts, and Humanistic Discourse: Toward Automatic Extraction and Classification', *Literary and Linguistic Computing*, 29.3 (2014), 452–64

Thelwall, Mike, and Maria M. Delgado, 'Arts and Humanities Research Evaluation: No Metrics Please just Data', *Journal of Documentation*, 71.4 (2015), 817–33

Tkaczyk, Dominika, and others, 'CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature', *International Journal on Document Analysis and Recognition*, 18.4 (2015), 317–35

——, 'Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers', *Proceedings of ACM JCDL* (2018), pp. 1–10

Waltman, Ludo, 'A Review of the Literature on Citation Impact Indicators', *Journal of Informetrics*, 10.2 (2016), 365–91

Wu, Jlian, and others, 'Citeseerx: AI in a Digital Library Search Engine', *Innovative Applications of AI Conference* (2017), 2930–37

Zuccala, Alesia, and others, 'Can we Rank Scholarly Book Publishers? A Bibliometric Experiment with the Field of History', *Journal of the Association for Information Science and Technology*, 66.7 (2015), 1333–47