DHQ: Digital Humanities Quarterly
2016
Volume 10 Number 2

# Exploring Citation Networks to Study Intertextuality in Classics

Matteo Romanello <matteo_dot_romanello_at_gmail_dot_com>, Deutsches Archäologisches Institut, Berlin / École Polytechnique Fédérale de Lausanne

## Abstract

Referring constitutes such an essential scholarly activity across disciplines that it has been regarded by [Unsworth 2000] as one of the scholarly primitives. In Classics, in particular, the references to passages of the ancient texts — the so-called *canonical citations* (or *references*) — play a prominent role. The potential of these citations, however, has not been fully exploited to date, despite the attention that they have recently received in the field of Digital Humanities.

In this paper I discuss two aspects of making such citations computable. Firstly, I illustrate how they can be extracted from text by using Natural Language Processing techniques, especially Named Entity Recognition. Secondly, I discuss the creation of a three-level citation network to formalise the web of relations between texts that canonical references implicitly constitute. As I outline in the conclusive section of this paper, the possible uses of the extracted citation network include the development of search applications and recommender systems for bibliography; the enhancement of digital environments to read primary sources with links to related secondary literature; and the application of these network to the study of intertextuality and text reception.

## 1 Introduction

Over the last two centuries Classics scholars have developed sophisticated tools and strategies to find relevant information they need for their work. These tools are aimed at making resources more easily accessible and include indexes of cited passages, specialised concordances and extensive bibliographic reviews, both critical and analytical. The fact that they are manually curated and therefore highly accurate, is what makes these resources valuable but time consuming to produce. This constitutes also the main limit of these resources as they cannot cope with the sheer amount of materials made available by large-scale digital libraries. 1

The result of this situation is that, when it comes to finding relevant resources within digital archives such as JSTOR, classicists are usually left with search functionalities based on string matching algorithms. In order to be exhaustive, a query to retrieve all journal articles that discuss a given ancient work needs to contain all variant spelling and abbreviations of the work title in several languages. For example, an exhaustive search for publications on Virgil's *Georgics* would need to include variant spellings of the title such as *Georgica*, *Georgics*, *Géorgiques*, *Georgiche*, *Geórgicas* and abbreviations such as “*Georg.*” and the less common “*G.*”. However, since building manually similar queries is rather inconvenient, a more scalable approach is required in order to provide scholars with the means of finding resources that are relevant to their research within large-scale digital archives. 2

The workbench of the 21st century classicist ought to offer more advanced means of searching for bibliographic information: a search for “Georgics” should return records that mention the title of the work in any of its variant forms or that cite specific sections of the poem (e.g. “Verg. *G.* 1.34”, “Verg. *Georg.* III 481 ”, etc.). Furthermore, it should be possible to search for articles on both Vergil and Lucan or articles that cite a specific set of text passages (e.g. Verg. *Aen.* 12.942 *and* Lucan 1.244). Such a specialised search, deemed by [Crane, Seales, and Terras 2009] to be one of the essential components of a cyberinfrastructure for Classics, is likely to have a particular impact on areas of study such as intertextuality where citations to primary sources play a key role. 3

Previous studies in the field of Digital Humanities have almost exclusively focused on the hypertextual dimension of canonical citations. Issues that were tackled by these studies include how — and with what consequences — such citations can technically be transformed into links and what new functionalities can thereby be provided in a digital reading environment [McCarty 2002] [Smith 2009] [Romanello 2011] [Kalvesmaki 2014]. In this paper I turn my attention to two additional aspects of these citations. First, how can they be extracted automatically from text? Second, how can the web of relations that these references constitute be formalised as a network? To illustrate my approach to these two issues I use a sample of reviews taken from *L'Année Philologique* in which canonical references were semi-automatically annotated. 4

This paper is organised as follows. In the first part I describe how canonical citations can automatically be extracted by applying Natural Language Processing (NLP) techniques. In the second part I discuss the creation of a citation network starting from the automatically extracted canonical references. The importance of such a network lies in that it gives formal representation to the web of relationships between texts that these references implicitly already constitute. I conclude this paper by sketching out what are the applications and further uses that such a citation network enables. 5

# 2 Extracting Citations as a Computational Problem

A considerable amount of time in Digital Humanities research is spent in trying to give a formal, computational formulation to problems of interest to humanities scholars. Broadly speaking, this is done by translating the problems into computational terms, turning them into computable tasks and representing them by means of data models. This process involves adapting existing methods and tools developed in disciplines such as Computer Science or Physics to these new scenarios as well as developing new ones. This certainly holds true for the extraction of citations to classical texts that are found in modern publications, such as commentaries or journal articles. 6

The approach to this problem that I adopted and built upon was first suggested by G. Crane [Crane, Seales, and Terras 2009] and consists of treating citations as a special kind of named entities. There is a whole area of research in Computer Science called Named Entity Recognition that deals with the automatic extraction of mentions of, among others, people, places and organizations (i.e. named entities). Capturing named entities implies also identifying the relations that exist between them (relation extraction) and disambiguating entity mentions by means of unique identifiers (named entity disambiguation). 7

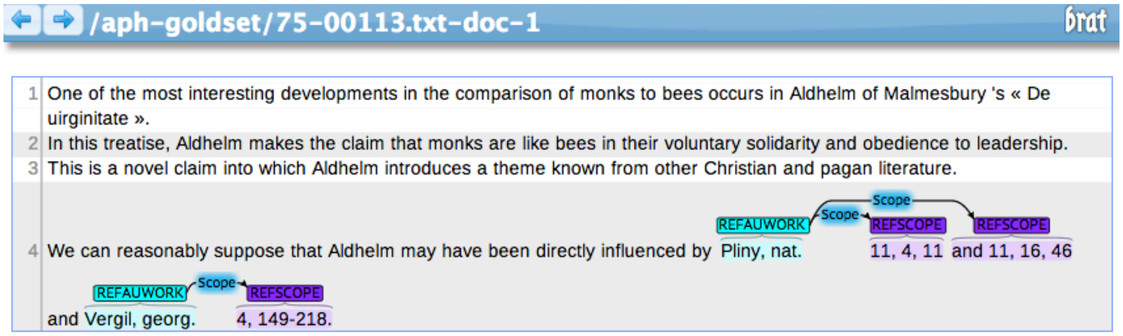
## 2.1 The Data: *L'Année Philologique*

The dataset that is considered in this paper is a sample of reviews drawn from *L'Année Philologique* (henceforth APh).[1] Since 1924 the APh has been reviewing annually what is published in every part of the world on a wide variety of topics in the Classics. Such a wide coverage makes the APh a fundamental bibliographic resource and the starting point for virtually any study of classical antiquity. However, this dataset covers only a very small portion of all the APh data: it contains slightly less than 30,000 words of text originating from 400 reviews that were drawn from a single volume out of the 80 volumes published to date. The volume used to create this dataset — APh vol. 75 — contains reviews of publications that appeared, or were reviewed, in 2004. 8

The main goal in the creation of this dataset was to train a piece of software to automatically extract citations and to evaluate the accuracy of the performed extraction.[2] The specific method that was used to create the dataset also determined the criteria for the selection of the reviews. This method is called Active Annotation and it aims to optimize the effort of manually annotating the data by selecting for inclusion in the training set the most informative documents [Romanello 2013]. In other words, the reviews that form the dataset were selected as they contain the citations whose automatic extraction proved to be most challenging for the software. 9

## 2.2 The Annotation Scheme

The first modelling choice that had to be made was to identify the named entities necessary to represent a range of canonical citations as wide as possible. Although it is true that such citations tend to have a rather homogeneous and somehow standardised format, the narrative within which they are situated leads to a wide range of possible variations 10 in their structure. What can vary substantially, for example, is the position within the sentence or the document that the components of a citation can take. The solution to this was to represent such citations as relations between the constituent components of a citation rather than as entities themselves.[3] For instance, when several passages of the same text are cited in a sequence the reference to the cited work ("Ath." and "Pol.") is given only once and then implied in all subsequent references: this can be captured by means of relations between entities (see Fig. 1).



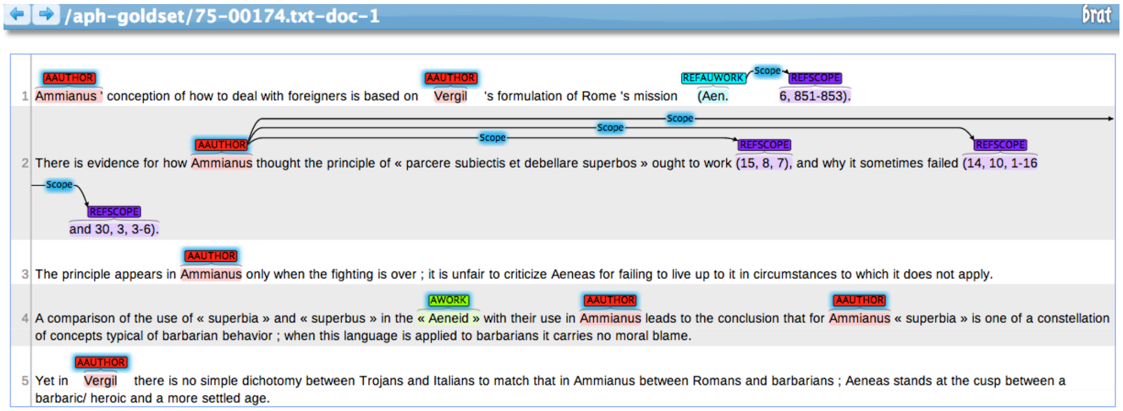
/aph-goldset/75-00113.txt-doc-1

1 One of the most interesting developments in the comparison of monks to bees occurs in Aldhelm of Malmesbury 's « De uirginitate ».
2 In this treatise, Aldhelm makes the claim that monks are like bees in their voluntary solidarity and obedience to leadership.
3 This is a novel claim into which Aldhelm introduces a theme known from other Christian and pagan literature.
4 We can reasonably suppose that Aldhelm may have been directly influenced by Pliny, nat. 11, 4, 11 and 11, 16, 46 and Vergil, georg. 4, 149-218.

**Figure 1.** An example of annotated APh document (75-0113) visualised in Brat: the highlighted portions of text indicate named entities, while the arrows represent the relations existing between them.

11 The example in Fig. 1 introduces the first two entity types: `REFAUWORK` and `REFSCOPE`. The former aims to capture the string indicating the text being cited — "Pliny, nat." and "Vergil georg." referring respectively to Plinius' *Naturalis Historia* and to Vergil's *Georgica* — whereas the latter captures the scope of the reference, that is the indication of which text passage is being cited ("11,4,11", "11,16,46" and "4,149-218").

12 In addition to `REFAUWORK` and `REFSCOPE`, the annotation scheme contains two other entities that capture respectively the name of an ancient author (`AAUTHOR`) and the title of an ancient work (`AWORK`) as shown in Fig. 2.



/aph-goldset/75-00174.txt-doc-1

1 Ammianus ' conception of how to deal with foreigners is based on Vergil 's formulation of Rome 's mission (Aen. 6, 851-853).
2 There is evidence for how Ammianus thought the principle of « parcere subiectis et debellare superbos » ought to work (15, 8, 7), and why it sometimes failed (14, 10, 1-16 and 30, 3, 3-6).
3 The principle appears in Ammianus only when the fighting is over ; it is unfair to criticize Aeneas for failing to live up to it in circumstances to which it does not apply.
4 A comparison of the use of « superbia » and « superbus » in the « Aeneid » with their use in Ammianus leads to the conclusion that for Ammianus « superbia » is one of a constellation of concepts typical of barbarian behavior ; when this language is applied to barbarians it carries no moral blame.
5 Yet in Vergil there is no simple dichotomy between Trojans and Italians to match that in Ammianus between Romans and barbarians ; Aeneas stands at the cusp between a barbaric/ heroic and a more settled age.

**Figure 2.** Representing canonical references as relations between named entities allows us to capture the discursive reference (line 2) "Ammianus [. . . ] (15, 8, 7) , and [...] (14, 10, 1-16 and 30, 3, 3-5)", which is made up of non consecutive tokens.

13 Although only the first two of the entities above capture the citation itself, the others are worth extracting as they may become useful when attempting to disambiguate the extracted citations. Let us consider the following example (named entities are highlighted in bold):

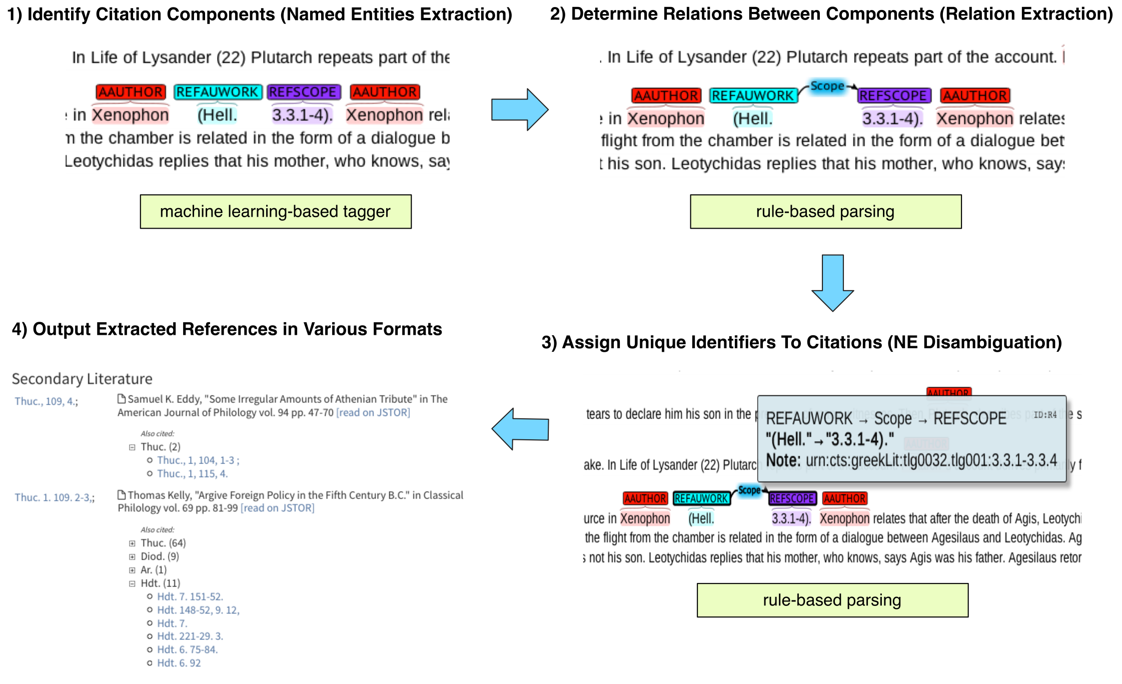
> [...] sind auch bei **Calpurnius Siculus (4.137ff.)**, **Sidonius Apollinaris (Carm. 5 und 7)** und **Ausonius (Epist. 17)** Anspielungen auf die « Apocolocyntosis » festzustellen.

Since there exist dozens of works titled *Epistulae* that can be referred to by the abbreviation "Epist." — the same applies also, for instance, to collections of *carmina* — it is almost impossible for the system to guess the correct disambiguation unless the neighbouring mention of the author Ausonius is captured.

In addition to these four named entities the annotation scheme includes a relation that captures the citation itself, named `scope`. A citation is defined as a relation existing between any two entities, where one must be the indication of the citation's scope (i.e. `REFSCOPE`) while the other can be any of the other entities (i.e. `AAUTHOR`, `AWORK` and `REFAUWORK`). 14

## 2.3 The Extraction Pipeline

How do we go from a plain text input to an output text that is annotated according to the scheme discussed above? This is done by a sequence of steps that form an extraction pipeline, each of them addressing a separate layer of annotation (see Figure 3).[4] The annotations were first produced automatically by running the data through the process described below and then corrected manually by two annotators. This semi-automatic process led to the creation of a dataset suitable to assess the accuracy of the automatic extraction of citations — in NLP jargon such a dataset is called *gold standard*. 15


1) Identify Citation Components (Named Entities Extraction)
In Life of Lysander (22) Plutarch repeats part of the
AAUTHOR
REFAUWORK
REFSCOPE
AAUTHOR
in Xenophon (Hell. 3.3.1-4). Xenophon rel
n the chamber is related in the form of a dialogue b
Leotychidas replies that his mother, who knows, say
machine learning-based tagger
2) Determine Relations Between Components (Relation Extraction)
. In Life of Lysander (22) Plutarch repeats part of the account.
AAUTHOR
REFAUWORK
Scope
REFSCOPE
AAUTHOR
in Xenophon (Hell. 3.3.1-4). Xenophon relates
flight from the chamber is related in the form of a dialogue bet
his son. Leotychidas replies that his mother, who knows, says
rule-based parsing
3) Assign Unique Identifiers To Citations (NE Disambiguation)
tears to declare him his son in the
ake. In Life of Lysander (22) Plutarc
AAUTHOR
REFAUWORK
Scope
REFSCOPE
AAUTHOR
urce in Xenophon (Hell. 3.3.1-4). Xenophon relates that after the death of Agis, Leotychi
the flight from the chamber is related in the form of a dialogue between Agesilaus and Leotychidas. Ag
not his son. Leotychidas replies that his mother, who knows, says Agis was his father. Agesilaus retor
rule-based parsing
4) Output Extracted References in Various Formats

REFAUWORK → Scope → REFSCOPE ID:R4
"(Hell."→"3.3.1-4)."
**Note:** urn:cts:greekLit:tlg0032.tlg001:3.3.1-3.3.4

Secondary Literature

Thuc., 109, 4.; Samuel K. Eddy, "Some Irregular Amounts of Athenian Tribute" in The American Journal of Philology vol. 94 pp. 47-70 [read on JSTOR]

*Also cited:*
⊟ Thuc. (2)
  ○ Thuc., 1, 104, 1-3 ;
  ○ Thuc., 1, 115, 4.

Thuc. 1. 109. 2-3,; Thomas Kelly, "Argive Foreign Policy in the Fifth Century B.C." in Classical Philology vol. 69 pp. 81-99 [read on JSTOR]

*Also cited:*
⊞ Thuc. (64)
⊞ Diod. (9)
⊞ Ar. (1)
⊟ Hdt. (11)
  ○ Hdt. 7. 151-52.
  ○ Hdt. 148-52, 9. 12,
  ○ Hdt. 7.
  ○ Hdt. 221-29. 3.
  ○ Hdt. 6. 75-84.
  ○ Hdt. 6. 92

Annotation Environment: http://brat.nlplab.org/

**Figure 3.** The various steps of the citation extraction pipeline.

The first step is the extraction of named entities from each document in the corpus. In order to do so a machine learning-based approach is employed, meaning that a statistical model is trained to predict, for each token (i.e. word) in the text, which label is to be assigned. During the training phase the model learns from the previously annotated data which features characterise tokens that are annotated with a given label, where each label corresponds to a named entity. Once trained, the model is then able to predict with some degree of accuracy the most likely labels for an unseen input sequence — i.e. a sequence that is not already contained in the training set. 16

The second step is the extraction of relations between named entities: as noted above, currently only the `scope` relation is considered. In the current implementation this is performed by using a rule-based approach: as opposed to the machine learning approach where the model learns how to perform a specific task based on a training set, in the rule-based approach a set of rules is defined based on some observations of the data. These rules take into account the position of named entities within the sentence as well as their position within the broader context of the document itself, as relations between entities may span across sentences. 17

In the third (and final) step the extracted named entities and relations are disambiguated, that is they are assigned a unique identifier. The identifiers of choice are Uniform Resource Names (URNs) that comply with the syntax specified by 18

Canonical Text Services (CTS) protocol [Smith 2010] [Smith and Blackwell 2012] [Smith 2009]. The CTS is a network protocol that was developed in the framework of the Homer Multitext project[5] and was designed to provide access to texts in a way that is consistent with how scholars have been working with such texts for centuries. In the case of canonical texts this means replicating in a digital environment what canonical citations have allowed scholars to already do in print, that is to create references to texts that are fine-grained and at the same time independent from any specific version (i.e. edition) of a text.

CTS URNs are used within the annotated data to identify unambiguously authors, works and even specific text passages: for example, the CTS URNs for Vergil, the *Aeneid* and the passage "*Aen.* 6.851-853" are respectively `urn:cts:latinLit:phi0690`, `urn:cts:latinLit:phi0690.phi003` and `urn:cts:latinLit:phi0690.phi003:6.851-6.853` (see Fig. 1). A more challenging example of disambiguation is provided by mentions of author names such as "Aristophanes" that can refer either to the Alexandrian grammar or to the comic playwright: in similar cases the broader context of the document needs to be considered in order to decide which author is being referred to, and thus which CTS URN is to be assigned to the entity. 19

Moreover, since a CTS URN encodes the scope of a notation in a normalised format in order for it to be machine readable, the citation needs to not only be disambiguated but also normalised: in the example above the scope "6.851-853" is normalised into "6.851-6.853". Similarly, the notation "6.851 s." — meaning book 6, line 851 and the following — needs to be made explicit and transformed into "6.851-6.852". The normalisation of citation scopes is also necessary because there are multiple ways of expressing the same citation. The citation scope "11.4.11", for instance, can be written also as "11,4,11" or "XI 4,11". 20

# 3 Texts Through the Lens of a Network

The extraction pipeline that was just discussed is the first step towards making fully explicit and computable the web of relations that canonical references implicitly constitute. The second step, which is discussed in this section, consists of transforming the extracted entities and relations into a formal network. This process implies decisions on, for example, which entity types will become nodes of the network and on the directionality of the connections between nodes (i.e. edges).[6] The decisions that are taken while creating the network are of crucial importance. Indeed, they shape the meaning of the network representation itself and determine what algorithms can be used for its exploration and analysis [Weingart 2012]. 21

Research in the field of citation network analysis has been focussing mainly on networks representing citations between modern publications (i.e. secondary sources). Such networks are used primarily to quantify the impact of publications by looking at the number of citations received or consider citation and co-authorship networks in order to analyse the structure and evolution of academic disciplines or their publishing and citing behaviours.[7] On the contrary, the study of citation networks that consist of references between primary and secondary sources, such as those discussed in this paper, remains a largely unexplored area of research. This fact may seem paradoxical given that roughly half of the citations contained within humanities publications refer to primary sources [Wiberley 2009, 2199]. A notable example from this area is the work by [Murai and Tokosumi 2005] and [Murai et al. 2008]. They have focussed in particular on canonical references to the Bible that are found within theological writings. Their analysis of the co-citation network of these references — i.e. which text passages of the Bible are cited in relation to each other — highlighted different conceptualisations of Christian dogma. 22

## 3.1 Citation Networks and the Study of Classical Texts

Networks of citations between modern publications represent "networks of relatedness of subject matter" [Newman 2010, 68]. The assumption underlying the use and analysis of these networks is that publications sharing a common subset of bibliographic references are also related to each other. But what is the meaning of a citation network created from canonical references? And what is its relevance for the study of classical texts? 23

Canonical references can be seen as *traces* that scholars leave in their publications in the form of citations. These 24 traces provide an indicator of what authors, works and text passages were studied — i.e. cited — and, at the same time, reflect how they were studied in relation to one another. Thus, the relatedness expressed by a citation network created from such references is two-fold. First, it is a relatedness between publications based on the primary sources they cite. Second, it is a relatedness between ancient authors or between ancient works that derives from how they are cited within modern publications.

Such a citation network lends itself to various uses. It can be used for information retrieval purposes in order to allow scholars to find publications that cite a specific set of text passages. Scholars with an interest in intertextuality would benefit most from such a means of searching for bibliographic information. In fact, the relations between texts that intertextuality investigates, such as allusions and other kinds of intertextual parallels, are indicated within modern publications by means of canonical references.[8] Once they are captured and formalised as a network, it becomes possible to search for publications that discuss a specific set of intertextual parallels. 25

Moreover, such a citation network can be used for quantitative studies on the reception of classical texts. The number of times a given author or text passage is cited can be taken as a proxy of the attention it received from scholars. If this citation network is extracted from publications covering a wider temporal span, it becomes possible not only to track variations in the *popularity* of ancient authors, expressed by the number of citations received by a given author, but also to observe how the sets of authors and works that are studied in relation to one another change over time. 26

## 3.2 A Three-Level Citation Network

The most challenging aspect of representing canonical citations as a formal network is how to preserve the multiplicity of hierarchical levels that such citations embody. For example, the reference Verg., *Aen.* 6.851-853 can be regarded as: a) a reference to Vergil; b) a reference to the *Aeneid* and c) a reference to lines 851-853 of book six of Vergil's *Aeneid*. Which level is to be taken into account depends from the context in which the citation network is used. When studying intertextuality, for example, one may want to consider intertextual relationships at the global level, that is between the entire body of works of two (or more) authors, or at a local level, meaning relationships between two (or more) specific works or even single text passages. 27

The approach I have taken to tackle this issue, inspired by a similar approach developed by [Schich and Coscia 2011], is to create a three-level network that allows us to look at the same citation data at different levels of abstraction, namely macro-, meso- and micro-level. Similar to how a lens works, these networks make it possible to produce a number of views on the citation data with an increasing degree of granularity and specificity. Each of these networks addresses a specific hierarchical level of analysis and is therefore likely to be useful to analyse and observe only certain phenomena. 28

These networks are all two-mode (or bipartite) and directed networks. Two-mode means that there are two types (or modes) of nodes in the network and that, by definition, edges can exist only between nodes with different modes. The definition of types, as I explain below, varies depending on the network level being considered. Moreover, since citations themselves have a directionality, that is from the citing document to the cited one, all three networks are *directed*, meaning that the citations are represented as edges going from the citing node to the cited one. 29

The network visualisations that follow were created from the manually corrected subset of the APh data, which consists of 366 documents — i.e. APh reviews — for a total of approximately 25,000 tokens and 850 canonical references. These visualisations use a force-layout algorithm to position the nodes on the canvas. As its name suggests, this algorithm works by applying different forces to each node in the network, namely repulsion, gravity and attraction. All nodes push each other away (repulsion), whilst connected nodes are pulled toward each other (attraction). Simultaneously, gravity pushes all nodes towards the center so as to oppose the repulsion and prevent the nodes from being pushed out of sight. The final configuration of the nodes results from the interplay of these three forces after several iterations of the algorithm. As a result, nodes that are highly connected with each other tend to remain in the middle of the canvas, whereas less connected nodes are pushed towards the periphery. 30

| Level | Modes | No. of nodes | No. of edges | Density |
|---|---|---|---|---|
| Macro | Document; ancient author | 396 | 404 | 0.0026 |
| Meso | Document; ancient work | 393 | 332 | 0.0022 |
| Micro | Document; text passage | 2,340 | 2,305 | 0.0004 |

**Table 1.** Basic statistics for each level of the citation network. The network was created using the manually corrected subset of APh reviews.

### 3.2.1 Macro-level Network

The macro-level network offers the most abstract view on the data and aims to provide a high-level perspective on the citations that are contained in a set of documents. Figure 4 shows a visualisation of the macro-level network extracted from the APh data, while some basic statistics on the size of the network are provided in Table 1. 31

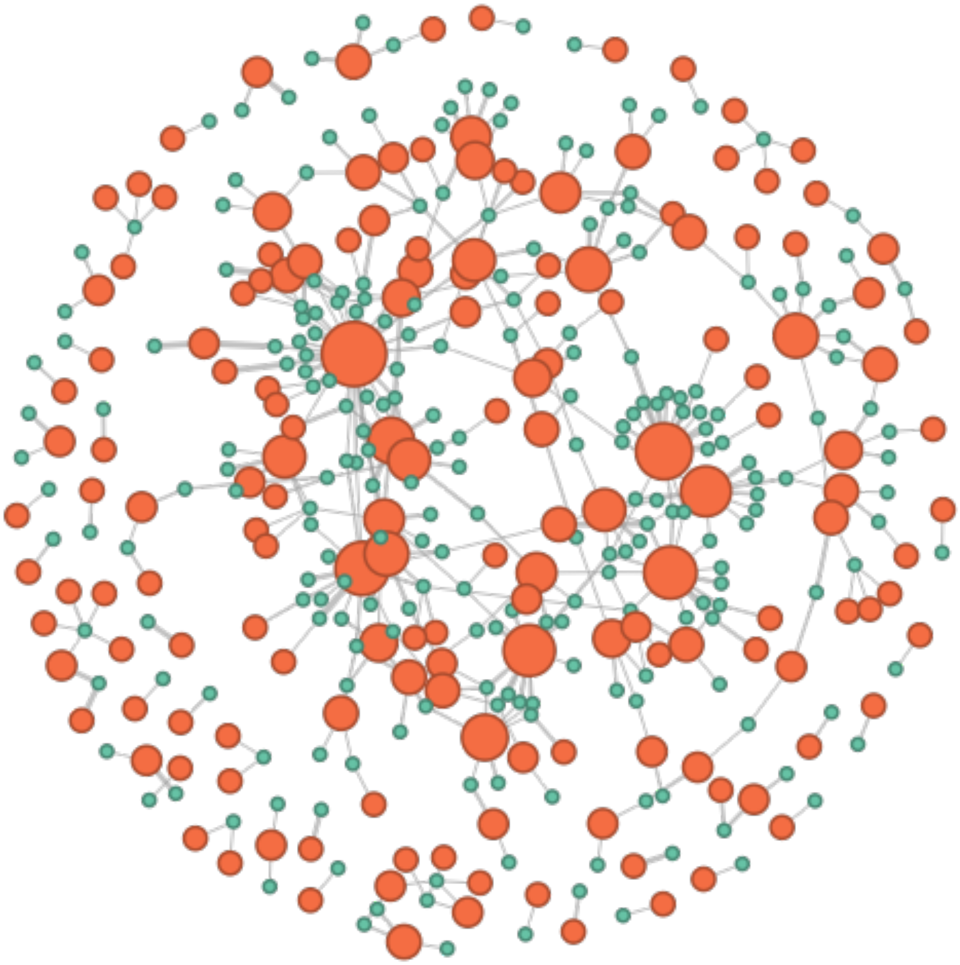


**Figure 4.** A visualisation of the macro-level citation network extracted from the APh data. The green nodes represent APh abstracts whilst the red nodes represent ancient authors. Although the directionality of the edges is not shown, the network is directed. The size of the nodes is proportional to their indegree (i.e. number of incoming citations).

Such a network is created by treating each canonical reference as a reference to the cited author while leaving aside the more detailed information about which work and specific text passage are cited. For example, the references “Pliny, *nat.* 11, 4, 11” and “Vergil, *georg.* 4, 149–218” contained in the document APh 75–00113 are treated as references to the cited authors — Pliny and Vergil. 32

This network is bipartite as there are two modes of nodes — APh documents and ancient authors — and there are no edges between nodes with the same mode. It is worth noting that an edge in this network can have two meanings: it can mean that a given author is explicitly cited but it can also mean that the author is simply mentioned in the text. In fact, as was described above, mentions of authors and works are extracted in addition to canonical references. Although it is desirable to capture both cases, it is also important for the meaning of the resulting network to be able to distinguish them. 33

Moreover, this two-mode, directed network can be projected into a one-mode undirected network where the nodes represent ancient authors. In this projection two authors are connected by an edge when they are cited by the same document. Such a projected network could be used in order to study to what extent the sets of authors that are studied and discussed in relation to one another change over time. 34

### 3.2.2 Meso-level Network

The meso-level network shown in Figure 5 offers a more detailed view of the data while maintaining some degree of abstraction compared to the micro-level. Canonical references are not treated as references to the cited author — as it is done at the macro-level — but to the cited work. For instance, the references “Pliny, *nat.* 11, 4, 11” and “Vergil, *georg.* 4, 149–218” of the example above are “compressed” respectively into a reference to Pliny’s *Naturalis Historia* and to *Vergil’s Georgics*. 35

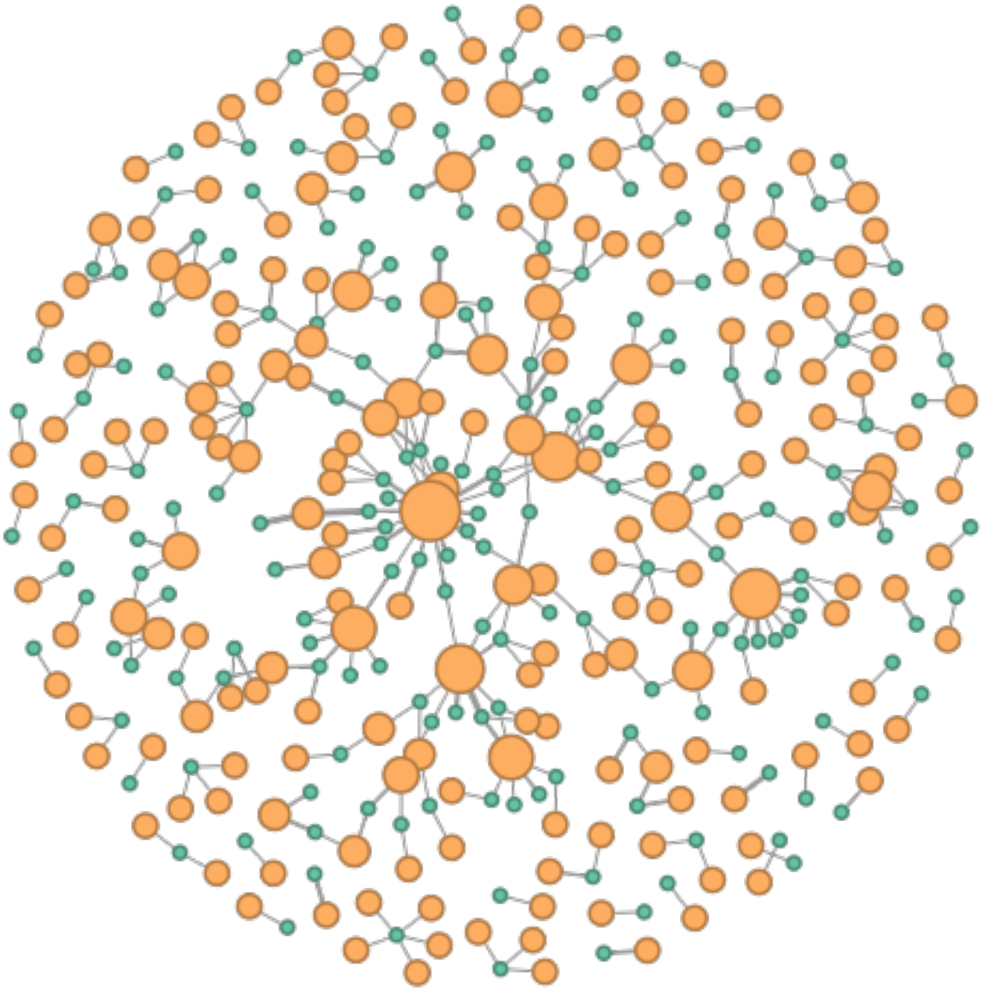


**Figure 5.** A visualisation of the meso-level citation network extracted from the APh data. The green nodes represent APh abstracts whilst the orange nodes represent ancient works.

The meso-level network shares the same properties as the macro-level network. Indeed, it is bipartite as it consists of 36 two types of nodes, documents and ancient works. Moreover, the edges are directed and, similar to the macro-level network, they can represent both mentions of titles of works and explicit references to specific sections of the work. Similarly, this network can be transformed into a one-mode undirected network where a relation between two ancient works is established whenever they are cited by the same APh document.

### 3.2.3 Micro-level Network

37 The highest degree of specificity and granularity is reached in the micro-level network (Figure 6). In this network each cited text passage is represented by a distinct node. References that point to a range of passages are expanded when creating this network: representing the reference “Vergil, *georg.* 4, 149–218”, for example, leads to creating additional nodes representing the lines comprised within the range 149-218. Performing this operation, which considerably increases the total number of nodes, has the advantage of making explicit references that are left implicit in the notation with which canonical references are usually expressed.

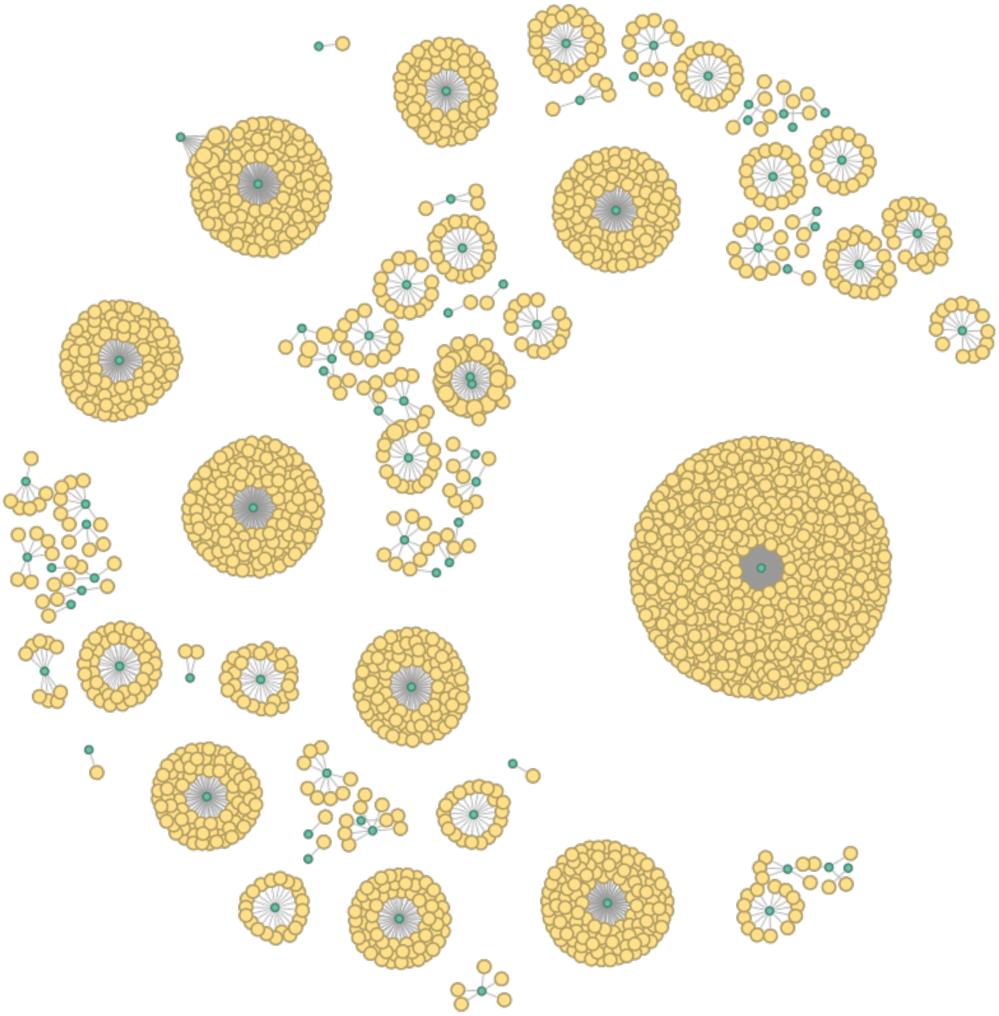


**Figure 6.** A visualisation of the micro-level citation network extracted from the APh data. The green nodes represent APh abstracts and the yellow nodes represent text passages.

38 The low degree of density characterising this network is what makes it most useful from an information retrieval point of view. In fact, searching this network makes it possible to retrieve documents that cite the very same set of text passages. It can be argued that such granular searches are already possible by using indexes of cited passages. However, since the networks on which the search is based are extracted automatically from text, it becomes possible to search through large-scale archives as if an index of all the text passages cited by the documents contained in these archives had been compiled.

Similar to the two previously examined levels, a further one-mode network can be projected from this micro-level network. In the resulting document-document network, two documents are connected when they share references to the same set of text passages. Such a network can be exploited in order to identify clusters of publications that are likely to be highly related to one another as they are concerned with the same primary sources.[9] 39

## 4 Conclusions and Further Work

In this paper I presented an approach to creating citation networks by automatically extracting canonical references to classical texts from modern publications. A relatively small dataset consisting of reviews drawn from *L'Année Philologique* was used to exemplify the extraction of such references as well as the creation of a three-level citation network that represents them. However, only the mining of these references on a larger scale may realise the full potential of this approach both in terms of user applications and of data analysis.[10] 40

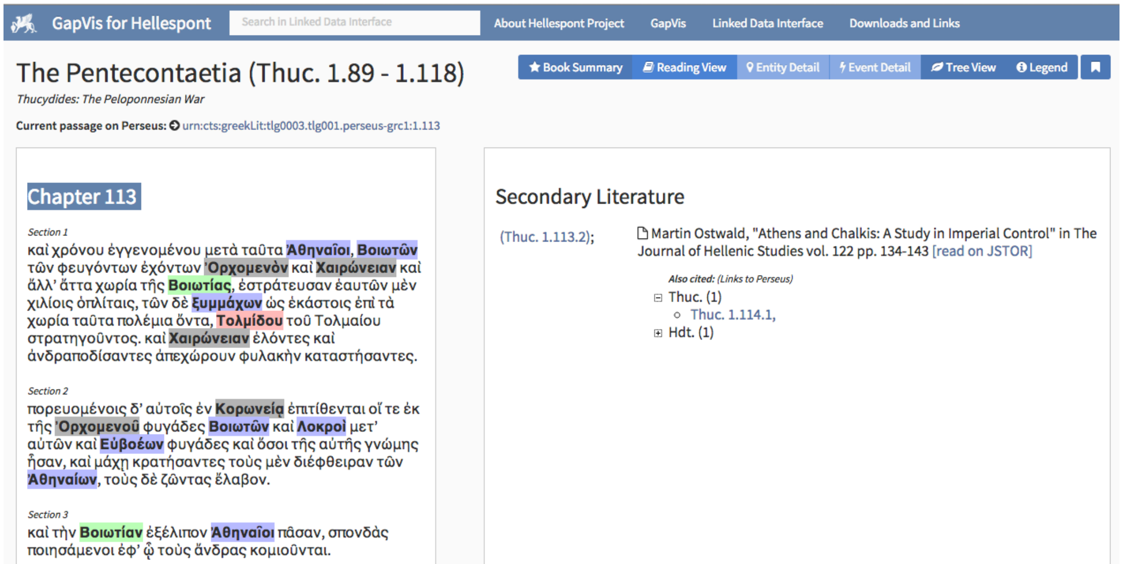
Several kinds of user applications could be developed building upon the citation networks that were described above. These applications include: 41

1. search applications;
2. recommender systems for bibliography;
3. enhanced reading environments.

Search applications would allow users to explore collections of publications using citations to primary sources as a key entry point to bibliographic information. In addition to searching by cited author and work, users would be able to retrieve documents that cite a specific text passage. This functionality is already provided, albeit on a smaller scale, by indexes of cited passages which constitute a scholarly resource of essential importance. Moreover, the fact that the relationships between texts are formalised as a network allows for using the graph — in addition to the hierarchical index — as a visual metaphor when designing the user interface for such a search application. In fact, the graph seems an apt way of representing visually and making browsable the connections between resources that are created by canonical references. 42

Recommender systems for bibliography — an increasingly common feature of digital libraries and reference management systems — often rely on the references contained in a given article to suggest related publications to the reader. While these systems take into consideration only references to other modern publications, the approach I described allows us to develop similar applications that leverage instead the references to classical texts. The cited primary sources become, in other words, the criterion to determine the relatedness between publications. The three-level network presented in the last section — and especially the document-document networks that can be projected at any level of the network — provide the citation data to which clustering algorithms could be applied in order to extract clusters of related publications. 43

Finally, canonical citations extracted from journal articles and other secondary sources can be employed within digital reading environments for primary sources so as to contextualise the text passage being read. This use of the citation data was explored within the Hellespont project [Romanello and Thomas 2012]. One of the outcomes of this project is an online environment to read the text of Thucydides' “Pentecontaetia” (Thuc. 1.89-118).[11] Such a reading environment presents to the user various layers of annotations with which the Greek text has been enriched. Additionally, a dedicated panel of the interface displays a list of articles contained in JSTOR that refer to the passage currently being read (see fig. 7).[12] To make this possible, canonical references to the “Pentecontaetia” were automatically mined from a sample of classics articles from JSTOR. 44



The Pentecontaetia (Thuc. 1.89 - 1.118)

*Thucydides: The Peloponnesian War*

**Current passage on Perseus:** urn:cts:greekLit:tlg0003.tlg001.perseus-grc1:1.113

Chapter 113

*Section 1*

καὶ χρόνου ἐγγενομένου μετὰ ταῦτα **Ἀθηναῖοι**, **Βοιωτῶν** τῶν φευγόντων ἐχόντων **Ὀρχομενὸν** καὶ **Χαιρώνειαν** καὶ ἄλλ᾽ ἄττα χωρία τῆς **Βοιωτίας**, ἐστράτευσαν ἑαυτῶν μὲν χιλίοις ὁπλίταις, τῶν δὲ **ξυμμάχων** ὡς ἑκάστοις ἐπὶ τὰ χωρία ταῦτα πολέμια ὄντα, **Τολμίδου** τοῦ Τολμαίου στρατηγοῦντος. καὶ **Χαιρώνειαν** ἑλόντες καὶ ἀνδραποδίσαντες ἀπεχώρουν φυλακὴν καταστήσαντες.

*Section 2*

πορευομένοις δ᾽ αὐτοῖς ἐν **Κορωνείᾳ** ἐπιτίθενται οἵ τε ἐκ τῆς **Ὀρχομενοῦ** φυγάδες **Βοιωτῶν** καὶ **Λοκροὶ** μετ᾽ αὐτῶν καὶ **Εὐβοέων** φυγάδες καὶ ὅσοι τῆς αὐτῆς γνώμης ἦσαν, καὶ μάχῃ κρατήσαντες τοὺς μὲν διέφθειραν τῶν **Ἀθηναίων**, τοὺς δὲ ζῶντας ἔλαβον.

*Section 3*

καὶ τὴν **Βοιωτίαν** ἐξέλιπον **Ἀθηναῖοι** πᾶσαν, σπονδὰς ποιησάμενοι ἐφ᾽ ᾧ τοὺς ἄνδρας κομιοῦνται.

Secondary Literature

(Thuc. 1.113.2); Martin Ostwald, "Athens and Chalkis: A Study in Imperial Control" in The Journal of Hellenic Studies vol. 122 pp. 134-143 [read on JSTOR]

*Also cited: (Links to Perseus)*

- Thuc. (1)
  - Thuc. 1.114.1,
- Hdt. (1)

**Figure 7.** The secondary literature view in the Hellespont reading environment. The Greek text of the passage in focus is displayed in the panel on the left, while the JSTOR articles related to this passage are shown on the right.

In addition to enabling the possible uses outlined above, the research presented here opens up new areas for further research. A first area concerns the design of a user interface that allows classicists to explore in an intuitive way collections of publications through this three-level citation network. Such an interface should enable the user to move back and forth between the different levels of analysis and to follow chains of citations within the network. A second area of research is constituted by the longitudinal study of this network, which looks at how the network evolves over time. This aspect was not considered in this paper as the APh reviews in the corpus cover only articles that were published or reviewed within a single year. On the contrary, a resource such as JSTOR would be ideally suited for this kind of analysis as it contains thousands of articles spanning across more than two centuries. Having at hand citation data covering a wider period of time has the potential to enable new approaches to the study of text reception in Classics. It will become possible, for example, to observe trends in the way ancient authors, works and even single text passages were objects of attention by scholars over time. 45

## Acknowledgments

Preliminary versions of this paper were presented at the Digital Classicist Association conference held at the University of Buffalo in 2013 and at the workshop “To Mine and to Tie: Text Mining and Network Analysis for Historians” held at the Ruhr University Bochum in 2015. I would like to thank the respective organisers of these events, Neil Coffee and Frederik Elwert, for having given me the chance to present and discuss my research. I am particularly grateful to Eric Rebillard and Adam Chandler for their help in getting access to and annotating the *L’Année Philologique* data and to Tom Brughmans, whose work on archaeological networks has inspired and influenced this research. 46

## Notes

[1] The annotated corpus is available under a Creative Commons licence at http://dx.doi.org/10.5281/zenodo.12762. For further technical details on how the corpus was created see [Romanello 2013].

[2] It should be noted that citations to primary sources in the online version of the APh are manually encoded as OpenURLs. These links can be resolved via the Classical Works Knowledge Base (CWKB) service available at http://cwkb.org/resservice.

[3] The former solution was the one adopted in [Romanello, Boschetti, and Crane 2009] but further refinements to the annotation scheme led to the development of the latter solution [Romanello 2013].

[4] The code I developed to implement this pipeline can be found at http://dx.doi.org/10.5281/zenodo.10886.

[5] The Homer Multitext Project, http://www.homermultitext.org/.

[6] For the formal definitions of the network science terms used in this paper the reader can refer to the glossary contained in [Collar et al. 2015]. The glossary is also available online at https://archaeologicalnetworks.wordpress.com/resources/#glossary.

[7] For an introduction to citation networks see [Newman 2010, 67–72] and [Radicchi, Fortunato, and Vespignani 2012]. [Brughmans 2012] provides an interesting example of how citation network analysis can be applied to literature in the field of Archaeology.

[8] Following [Coffee et al. 2012] I use the word *intertextuality* to mean a wide range of relationships between texts. For an extensive annotated bibliography on the various aspects and literary traditions that the study of intertextuality encompasses see [Coffee 2013].

[9] It should be noted that the projected document-document network resembles closely a bibliographic coupling network, which can be extracted from a citation network (i.e. two publications are coupled when they cite a common third publication). For further details on bibliographic coupling networks see [Newman 2010, 115–118].

[10] This is an area that I am still researching as I am currently working on the extraction of canonical citations from the Classics journal articles contained in JSTOR.

[11] GapVis for Hellespont, http://gapvis.hellespont.dainst.org/#book/1/read/89/.

[12] A similar functionality is provided by the “criticism” facet of Segetes, http://segetes.io/. For an example of interfaces built using the Segetes framework see http://segetes.io/aeneid/.

## Works Cited

**Brughmans 2012** Brughmans, Tom. 2012. “Thinking Through Networks: A Review of Formal Network Methods in Archaeology.” *Journal of Archaeological Method and Theory*. Springer US, 1–40. doi:10.1007/s10816-012-9133-8.

**Coffee 2013** Coffee, Neil. 2013. “Intertextuality in Latin Poetry.” Edited by Dee L. Clayman. doi:10.1093/obo/9780195389661-0113.

**Coffee et al. 2012** Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Roelant Ossewaarde, Christopher Forstall, and Sarah Jacobson. 2012. “Intertextuality in the Digital Age.” *Transactions of the American Philological Association* 142 (2): 383–422. doi:10.1353/apa.2012.0010.

**Collar et al. 2015** Collar, Anna, Fiona Coward, Tom Brughmans, and Barbara J Mills. 2015. “Networks in Archaeology: Phenomena, Abstraction, Representation.” *Journal of Archaeological Method and Theory* 22 (1): 1–32. doi:10.1007/s10816-014-9235-6.

**Crane, Seales, and Terras 2009** 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.

**Kalvesmaki 2014** Kalvesmaki, Joel. 2014. “Canonical References in Electronic Texts: Rationale and Best Practices.” *Digital Humanities Quarterly* 8 (2). http://www.digitalhumanities.org/dhq/vol/8/2/000181/000181.html.

**McCarty 2002** McCarty, Willard. 2002. “A Network with a Thousand Entrances: Commentary in an Electronic Age?” In *The Classical Commentary: Histories, Practices, Theory*, edited by Roy K Gibson and Christina Shuttleworth Kraus, 359–402. Mnemosyne Supplements.

**Murai and Tokosumi 2005** Murai, Hajime, and Akifumi Tokosumi. 2005. “A Network Analysis of Hermeneutic Documents Based on Bible Citations.” In *CogSci 2005*, edited by Bruno G. Bara, Lawrence Barsalou, and Monica Bucciarelli.

**Murai et al. 2008** Murai, Hajime, Akifumi Tokosumi, Takenobu Tokunaga, and Antonio Ortega. 2008. “Extracting concepts from religious knowledge resources and constructing classic analysis systems.” In *Large-Scale Knowledge Resources. Construction*, edited by Takenobu Tokunaga and Antonio Ortega, 4938:51–58. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. http://www.springerlink.com/index/u33485j7757x5ngh.pdf.

**Newman 2010** Newman, Mark. 2010. *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc.

**Radicchi, Fortunato, and Vespignani 2012** Radicchi, Filippo, Santo Fortunato, and Alessandro Vespignani. 2012. “Citation Networks.” In *Models of Science Dynamics*, edited by Andrea Scharnhorst, Katy Börner, and Peter van den Besselaar, 233–57. Understanding Complex Systems. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-23068-4_7.

**Romanello 2011** Romanello, Matteo. 2011. “New Value-Added Services for Electronic Journals in Classics.” *JLIS.it* 2 (1). doi:10.4403/jlis.it-4603.

**Romanello 2013** Romanello, Matteo. 2013. “Creating an Annotated Corpus for Extracting Canonical Citations from Classics-Related Texts by Using Active Annotation.” In *Computational Linguistics and Intelligent Text Processing. 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*, edited by Alexander Gelbukh, 1:60–76. Lecture Notes in Computer Science / Theoretical Computer Science and General Issues. Springer Berlin Heidelberg. doi:10.1007/978-3-642-37247-6_6.

**Romanello and Thomas 2012** Romanello, Matteo, and Agnes Thomas. 2012. “The World of Thucydides: From Texts to Artefacts and Back.” In *Revive the Past. Proceeding of the 39th Conference on Computer Applications and Quantitative Methods in Archaeology. Beijing, 12-16 April 2011*, edited by Mingquan Zhou, Iza Romanowska, Wu Zhongke, Xu Pengfei, and Philip Verhagen, 276–84. Amsterdam University Press. http://dare.uva.nl/document/358465.

**Romanello, Boschetti, and Crane 2009** Romanello, Matteo, Federico Boschetti, and Gregory Crane. 2009. “Citations in the digital library of classics: extracting canonical references by using conditional random fields.” In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, 80–87. NLPIR4DL ’09. Morristown, NJ, USA: Association for Computational Linguistics.

**Schich and Coscia 2011** Schich, Maximilian, and Michele Coscia. 2011. “Exploring Co-Occurrence on a Meso and Global Level Using Network Analysis and Rule Mining.” In *Proceedings of the Ninth Workshop on Mining and Learning with Graphs MLG 11*. ACM.

**Smith 2009** Smith, Neel. 2009. “Citation in Classical Studies.” *Digital Humanities Quarterly* 3 (1). http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html.

**Smith 2010** Smith, Neel. 2010. “Digital Infrastructure and the Homer Multitext Project.” In *Digital Research in the Study of Classical Antiquity*, edited by Gabriel Bodard and Simon Mahony, 121–37. Burlington, VT: Ashgate Publishing.

**Smith and Blackwell 2012** Smith, Neel, and Christopher Blackwell. 2012. “Homer Multitext Project: documentation. An overview of the CTS URN notation.” PhD thesis. http://www.homermultitext.org/hmt-doc/cite/cts-urn-overview.html.

**Unsworth 2000** Unsworth, John. 2000. “Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?” http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

**Weingart 2012** Weingart, Scott B. 2012. “Demystifying Networks, Parts I & II.” http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart/.

**Wiberley 2009** Wiberley Jr., Stephen E. 2009. “Humanities Literatures and Their Users.” http://hdl.handle.net/10027/7012.