# Citations in the Digital Library of Classics: Extracting Canonical References by Using Conditional Random Fields

**Matteo Romanello, Federico Boschetti, Gregory Crane**
The Perseus Project
Medford, MA, USA
matteo.romanello, federico.boschetti, gregory.crane{@tufts.edu}

## Abstract

Scholars of Classics cite ancient texts by using abridged citations called canonical references. In the scholarly digital library, canonical references create a complex textile of links between ancient and modern sources reflecting the deep hypertextual nature of texts in this field. This paper aims to demonstrate the suitability of Conditional Random Fields (CRF) for extracting this particular kind of reference from unstructured texts in order to enhance the capabilities of navigating and aggregating scholarly electronic resources. In particular, we developed a parser which recognizes word level n-grams of a text as being canonical references by using a CRF model trained with both positive and negative examples.

## 1 Introduction

In the field of Classics, canonical references are the traditional way established by scholars to cite primary sources within secondary sources. By primary sources we mean essentially the ancient texts that are the specific research object of Philology, whereas by secondary sources we indicate all the modern publications containing scholarly interpretations about those ancient texts. This specific characteristic strongly differentiates canonical references from the typical references we usually find within research papers.

Canonical references are used to shortly refer to the research object itself (in this case ancient texts) rather than to the existing literature about a certain topic, as happens with references to other secondary sources. Given this distinction, canonical references assume a role of primary importance as the main entry point to the information contained in scholarly digital libraries of Classics. To find a parallel with other research fields, the role played by those references is somewhat analogous with that played by protein names in the medical literature or by notations of chemical compounds in the field of Chemistry. As was recently shown by Doms and Schroeder (2005) protein names can be used to semantically index documents and thus to enhance the information retrieval from a digital library of texts, provided that they are properly organized by using an ontology or a controlled vocabulary. Moreover, by analyzing and indexing such references as if they were backlinks (Lester, 2007) from a secondary to a primary source, it is possible to provide quantitative data about the impact of an ancient author for research in a particular disciplinary field, or in relation to a limited corpus of texts (e.g., the papers published by scholarly journals in a given time interval).

In addition to serving as entry points to information, canonical references can also be thought of as a navigation apparatus that allows scholars to browse seamlessly through ancient texts and modern interpretations about them (Crane, 1987). For every scholar working on the ancient historiographer Herodotus, for instance, it would be extremely useful to be able to easily access all the secondary sources containing references to Herodotus' works.

Therefore, the ability to automatically identify canonical references within unstructured texts is a first and necessary step to provide the users of digital libraries of Classics with a more sophisticated way to access information and to navigate through the texts that are already available to scholars of other fields.

The volume of publicly available digitized books constituting what has been called the Million Book Library (Crane, 2006) has made it essential to develop automatic and scalable tools to automate the process of information extraction from electronic resources. Furthermore, the obso-



*Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009*, pages 80–87,
Suntec, Singapore, 7 August 2009. 

lescence time for publications is far longer in Classics than in other disciplines, meaning that typically the value of a publication does not decrease drastically after a certain time. As a result, scholars in Classics may be the most potential beneficiaries of the recent mass digitization initiatives, since they have already started with many materials out of copyright.

In this paper we describe how Conditional Random Fields (Lafferty et al., 2001), the state of the art model in automatic classification, can be suitably applied to provide a scalable solution to this problem.

## 2 Related work

Canonical references to primary sources can be explored from at least three different angles: 1) identification and extraction; 2) hypertextual navigation; 3) semantics.

The identification and extraction of bibliographic references from what we called secondary sources (i.e. monographs, commentaries, journal papers, etc.) is a well explored task for which effective tools already exist. Although the biggest efforts in this direction have been made in the scientific fields, those tools can also be suitably adapted to the field of Classics, since they are essentially based on machine learning techniques.

Several researchers recently focused on applying computational linguistics methods to automatically extract information from both Classical texts and modern texts about them, in order to support the above described needs of scalability. Gerlach and Crane (2008), and Kolak and Schilit (2008) considered the identification of citations within primary sources by analyzing the syntactic and morphological features of texts, while (Smith and Crane, 2001) dealt with the disambiguation of geographical names.

Looking at the problem of canonical references from the user point of view, a digital library of Classical texts such as the Perseus Digital Library[1]. already offers to the reader the ability to navigate from secondary sources to the primary sources they refer to, a process called reference linking. The identification of references and the attribution of semantics to them, however, was done manually, and the navigation is limited to resources contained in the same text collection. An analogous reference linking system was proposed by Romanello (2008) as a value added service that could be provided to readers of electronic journals by leveraging semantic encoded canonical references.

(Smith, 2009) provided an essential contribution to the research concerning the semantics of canonical references. The Canonical Text Services (CTS) protocol[2] was developed by Smith for Harvard's Center for Hellenic Studies; it is based on URNs and is aimed at providing a machine actionable equivalent to printed canonical references. This protocol allows us to translate those references into machine actionable URNs that can then be resolved through resolution services against a distributed digital library of texts. The innovative aspect of the CTS protocol consists of a loose coupling system by which the linking between primary and secondary sources can be realized. Instead of hard linking a canonical reference to just one electronic edition of a primary source, by embedding the CTS URNs inside (X)HTML pages, it becomes possible to link it to an open ended number of resources as shown by (Romanello, 2007).

## 3 Canonical Text References

Canonical references present unique characteristics when compared to bibliographic references to modern publications. First of all, they do not refer to physical facts of the referred work (such as publication date or page number), but refer rather to its logical and hierarchical structure. In addition, canonical references often provide additional information needed by the reader to resolve the reference. For example "Archestr. fr. 30.1 Olson-Sens" means line 1 of fragment 30 of the comic poet Archestratus in the edition published by S. D. Olson and A. Sens in 1994.

The specification of the edition according to which a source is cited is an important piece of information to be considered. Indeed, since the aim of Philology is to reconstruct for ancient works a text that is as close as possible to the original one (given that the original text may have been corrupted over centuries of manuscript tradition), editors and scholars often disagree substantially as to what readings and conjectures have to be included in the established text.

Although some well established sets of abbreviations exist, scholars' practice of citing primary

[1] http://www.perseus.tufts.edu/hopper/

[2] http://chs75.harvard.edu/projects/diginc/techpub/cts

sources may noticeably differ according to style preferences and the typographical needs of publishers, journals or research groups. Aeschylus' name might appear in the abridged forms "A., Aesch., Aeschyl.", and similarly a collection of fragments like Jacoby's *Die Fragmente der Grieschischen Historiker* may be abbreviated either as FrGrHist or FGrHist.

Moreover, some highly specialized branches of research exist within the field of Classics, such as those dedicated to Epic poetry or Tragedy, or even to a single author like Aeschylus or Homer. In those specialized branches a common tendency to use shorter references with a higher semantic density for the most cited authors can be observed. For example, in publications containing thousands of references to Homer's *Iliad* and *Odyssey*, references to these texts are often expressed with Greek letters indicating the book number along with the verse number (e.g., "α 1" stands for the first verse of the first book of Homer's *Odyssey*). Lowercase letters are used to refer to books of the *Odyssey*, whereas uppercase letters refer to the books of the *Iliad*, according to a practice developed in the IV century B.C. by scholars of the library at Alexandria.

In the actual practice of scholarly writing, canonical references can appear with slightly different figures according to the needs of narrative. Along with complete canonical references to a single text passage, expressed as either a single value or a range of values, other references can often be found that are missing one or more components that are normally present within canonical references, such as an indication of the author name, of the work title or of the editor name (e.g., "Hom. Od. 9.1, 9.2-3; Il 1.100"). This happens particularly in subsequent references to passages of the same work.

Those differences that can be observed about the appearance of canonical references require us to apply different processing strategies to each case. We focus on the task of automatically identifying complete references to primary sources. Once those references have been identified in the input document, we can find other anaphoric references by applying some scope-based parsing. Indeed, a canonical reference in the text constitutes the *reference scope* for subsequent text passage indications referring to the same work.

## 4 Methodology

Provided that scholars may use canonical references with different abbreviation or citation styles, it is nevertheless possible to identify within canonical references common patterns in terms of token features.

CRF is used to classify a token depending on its features and is suitable to identify those feature patterns (Culotta et al., 2006). During the training phase, the CRF model learns what features make it more likely for a token to belong to a given category.

Our starting assumption is that it is possible to determine if a sequence of tokens constitute a canonical reference by evaluating (looking at) the features of its tokens. Each token of a sequence is assigned a category on the basis of a fixed number of features. Those token categories are in turn used as features to classify the token sequence.

Starting from a dataset of canonical references and applying the above described criteria to assign features to the tokens, we obtain a training dataset where each canonical reference is reduced to a token by removing whitespaces, and it is a assigned as many as features as the category assigned to its tokens.

Finally, in order to classify token sequences as "references" or "non-references" each canonical reference is assigned a convenient label. The obtained set of labelled references is used to train a CRF model to identify canonical references within unstructured texts.

### 4.1 Feature Extraction and Token Categorization

For feature extraction phase, it was important to identify both inclusive and exclusive token features. Indeed, to extract canonical references with a high level of precision, we need to identify not only the characteristic features of tokens occurring within actual references but also those characteristic features for tokens occurring in sequences that we want to be classified as non-references.

Even though the features are quite similar to those used to identify modern bibliographic references (Isaac Councill and Kan, 2008), they were tuned to fit the specific needs of canonical references to primary sources. We decided to record a total of 9 features for each token, concerning the following aspects:

1. *Punctuation*: information about the punctuation concerning the presence of a final dot, hyphen, quotation marks and brackets (either single or paired), and marks used to divide and structure sequences (i.e. comma, colon and semicolon), which are particularly important for sequences of text passages.

2. *Orthographic Case*: the orthographic case of a token is an essential piece of information to be tracked. Author names when abbreviated still keep the initial as an uppercase letter, whereas collections of texts (such as collections of fragments) often present all uppercase or mixed case letters (e.g., "TrGrFr","CGF", "FHG", etc.).

3. *Stopwords*: given that the main language of the input document is passed as a parameter to the parser, we record in a separate feature information regarding whether a token is a stopword in the input document language. This feature is particularly important in determining more precisely the actual boundaries of a canonical reference within the text.

4. *Greek Words*: since we deal with Unicode UTF-8 text, we distinguish Greek letters and words. This allows us to identify more precisely those references that contain Greek text such as the above mentioned Homeric references or references to the ancient lexica (e.g., Harpocr., Lex. s.v. Παναθήναια) since they contain the lemma of the Greek word referred to, usually preceded by the abbreviation "s.v." (i.e. sub voce).

5. *Number*: Roman and Arabic numerals combined in several figures are frequently used to indicate the scope of a reference. Arabic numerals that are used to represent modern dates, however, are distinguished by using a heuristic (for example, consider the problem of a footnote mark which gets appended to a date). Nevertheless, sequences of both numbers and punctuation marks are assigned a specific value for this feature, since the scope of a reference is commonly expressed by dot and hyphen separated sequences such as "9.235-255".

6. *Dictionary matching*: two features are assigned if a token matches a dictionary entry. Three different dictionaries are used to verify if a token corresponds to a known canonical abbreviation (e.g. "Hom." for Homer or "Od." for *Odyssey*) or to another kind of abbreviation, namely the abbreviations used by philologists to shortly refer to pages, lines, verses, etc. ("p", "pp.", "v.", "vv.", "cfr", etc.) or to abbreviations used for modern journals. Abbreviations pertaining to the latter kind are likely to introduce some noise during the n-gram classification phase and thus are properly distinguished through a specific feature. During preliminary analysis we particularly observed that journal abbreviations were often confused with abbreviations for text collections since - as we noted above - they share the feature of having uppercase or mixed case letters.

7. *Fragment indication*: canonical references to fragments usually contain the indication "fr." (and "frr." for more than one). Therefore we expect tokens bearing this feature to occur almost exclusively within references to fragmentary texts.

We extract from the training dataset those unique patterns of these 9 token features that are likely to be found within canonical references. In order to ensure both the scalability and the extensibility of the suggested method to disciplinary fields other than Classics, we did not assign an identity feature to tokens or - in other words - the actual string content is not considered as a token feature. However, since this decision might decrease the overall precision of the system, we introduced some features to record whether the token string occurs in one or more controlled dictionaries (e.g., list of widely adopted abbreviations).

An analogous consideration is valid also for the dependency of the system from a specific language. Even though the approach is substantially language independent, the performances of our system in terms of precision were improved by using language specific lists of stopwords in order to identify the actual boundaries of a canonical reference within the text. Currently we support the most commonly used languages in the field of Classics (English, French, German, Italian, Spanish).

Finally, it is worth noting that the use of italics is a distinctive feature in particular for those tokens

that represent abbreviations of work titles. Since we are dealing with plain text input documents, however, and wish to keep the adopted approach as generalizable as possible, this feature has not been taken into account.

| **Token** | **Features** | | | | | | | | | **Cat.** |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | |
| Od. | ICP | FDT | NOD | OTH | OTH | OTH | CAB | OTH | OTH | 1_c50 |
| 9.216-535. | OTH | FDT | DSN | OTH | OTH | OTH | OTH | OTH | OTH | 2_c6 |

Table 1: Categorization of tokens of the reference "Od. 9.216-535" on the basis of their features.

| **Token** | **Features** | | **Cat.** |
|---|---|---|---|
| | F1 | F2 | |
| Od._9.216-535 | 1_c50 | 2_c6 | ref |

Table 2: Categorization of the reference of Tab. 1 by using token categories as its features.

| **Feature Label** | |
|---|---|
| F1 | Case |
| F2 | Punctuation Mark |
| F3 | Number |
| F4 | Greek Sequence |
| F5 | Stop Word |
| F6 | Paired Brackets |
| F7 | Contained in the 1st Dict. |
| F8 | Contained in the 2nd Dict. |
| F9 | Fragment Indication |
| **Feature Value** | |
| CAB | Canonical Abbreviation |
| DSN | Dot Separated Number Plus Range |
| FDT | Final Dot |
| ICP | Initial Cap |
| NOD | No Digit Sequence |
| OTH | Other |

Table 3: List of abbreviations used in Tab. 1, 2.

### 4.2 Positive and Negative Training

Since the main goal of our parser is to identify canonical references by isolating them from the surrounding context, both positive and negative training examples are needed. Indeed, provided two token sequences where the first contains just a canonical reference (e.g., "Od. 9.216-535") and the second additionally includes some tokens from the context phrase (e.g.,"Od. 9.216-535, cfr. p. 29."), without a negative training phrase both token sequences would have the same degree of similarity. When weighted by the CRF model the result would be that both sequences would share the same number of features with one of the references of the positive training. But since other sequences presenting features from both the positive and negative training were included in the training, and since such sequences were labelled as "non-references", the end result is that a token sequence with some tokens from a context phrase will be less similar to a pure canonical reference.

The first step of the training phase is the extraction of token features and the identification of unique patterns of token features. At this stage the processing units are the tokens of a reference. Given a dataset of canonical references, each reference is firstly tokenized and each token is then assigned 9 labels containing the values for the above described features (see Section 4.1). Note that in Tab. 1, 2 the labels and values of features are indicated by the abbreviations given in Tab. 3.

The observed combinations of feature values are then deduplicated and rearranged into unique categories that are used to classify each token (see Tab 1). These categories correspond to the uniques combinations of features assigned to tokens of references in the training dataset. Each category is defined by a name such as "c6" or "c50", where "c" simply stands for 'category'" and "6" or "50" are unique numeric identifiers. Besides, a numerical prefix corresponding to the position of the token inside the canonical reference is then added to the category name to form the identifier. Indeed, the position of each token in the sequence is in itself meaningful information, provided that indications of the reference scope (and other reference components as well) tend to occur at the end of the token sequence. What we obtain are category identifiers such as "1_c50" or "2_c6".

The second step is building the training dataset. At this stage each canonical reference is reduced to a single token which is assigned the label "ref" (i.e. reference) and which has as distinctive features the category identifiers assigned to its tokens (see Tab 2).

Finally, a such obtained dataset of labelled instances is used to train our CRF model by using the Java CRF implementation provided by the Mallet toolkit (McCallum, 2002).

### 4.3 Sequence Classification Process

The system we propose to identify canonical references in unstructured texts is basically a binary classifier. Indeed, it classifies as "reference" or "non-reference" a sequence of word level n-grams depending on the features of its tokens. However, in the training dataset the positive examples are manually grouped by typology and different labels (such as "ref1", "ref2" etc.) are assigned to canonical references pertaining to different types. This is done in order to avoid associating too many features to a single class and thus to maximize the difference in terms of features between sequence being references and non-references.

Since every token is assigned a certain number of features and finally a category, the likelihood for a token sequence to be a canonical reference can be determined on the basis of its similarity, in terms of token features, to the labelled references of a training set.

Once the input document is tokenized into single words, the n-grams are created by using a window of variable dimensions ranging from the minimum to the maximum length in terms of tokens that was observed for all the references in the training dataset. For example, provided that the shortest canonical reference in the training dataset is 2 tokens long and the longest is 7 tokens long, for each token are created 6 word level n-grams.

For the sake of performance, however, the number of n-grams to be created is determined for each token at parsing time. First of all a threshold value is passed to the parser as an option value. The threshold is compared to the weight value assigned by the CRF model to the probability of a token to be classified with a label, in our case "ref" or "noref". For each token, if the first n-gram is classified as not being a canonical reference the processing shifts to the next token, since we observed that if the first n-gram is classified as a non-reference the following n-grams of increasing width never contain a reference. If the examined n-gram is classified as reference, another of dimension n+1 is created: the parser passes on to process the next token only if the current n-gram is classified as a canonical reference with a likelihood value greater that that of the previous n-gram.

## 5 Training and Evaluation Criteria

The system is based on both a positive and a negative training.

The dataset for the positive training is built by labeling with the above explained criteria a starting set of approximatively 50 canonical references selected by an expert. The classifier trained with those positive examples is then applied to a random set of documents. Extracted candidate canonical references are scored by the CRF model by assigning to each sequence of n-grams a value representing the probability for the sequence to be a canonical reference.

The first one hundred errors with the highest score, due to the sharing of several features with the actual canonical references, are marked as non-references and added to the set of sequences to use for the negative training. The negative training is needed in order to precisely segment a canonical reference and to correctly classify those sequences that are most likely to be confused with actual canonical references, such as sequences only partially containing a canonical reference or bibliographic references. In particular, bibliographic references are misleading sequences since they have several features in common with canonical references, such as capitalized titles and page numbers.

The overall performances of the system on a random sample of 24 pages can be summarized by: precision=81.01%, recall=94.11%, accuracy=77.11%, F-score=0.8707. Analytical data are provided in Tab. 4. Although the evaluation was performed on pages drawn from a publication written in Italian, we expect to have analogous performances on texts written in each of the currently supported languages (English, French, German, Italian, Spanish) for the reasons described in Section 4.1.

The results are encouraging, however, and some further improvements could concern the recovery of tokens wrongly included in or excluded from the sequence identified by the parser.

## 6 Conclusion and Future Work

This paper has illustrated how the CRF model can be suitably applied to the task of extracting canonical references from unstructured texts by correctly classifying word level n-grams as references or non-references.

| Document # | Precision | Recall | Accuracy | F-Score |
|---|---|---|---|---|
| 40 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 41 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 55 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 57 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 62 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 64 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 67 | 25.00% | 25.00% | 25.00% | 0.2500 |
| 74 | 88.00% | 87.50% | 77.78% | 0.8800 |
| 77 | 45.00% | 90.00% | 42.86% | 0.6000 |
| 82 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 85 | 100.00% | 90.00% | 90.00% | 0.9474 |
| 88 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 90 | 92.31% | 92.31% | 85.71% | 0.4286 |
| 100 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 113 | 60.00% | 100.00% | 60.00% | 0.7500 |
| 117 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 134 | 100.00% | 75.00% | 75.00% | 0.8571 |
| 137 | 75.00% | 100.00% | 75.00% | 0.8571 |
| 144 | 67.00% | 100.00% | 67.00% | 0.8024 |
| 146 | 33.00% | 100.00% | 33.00% | 0.4511 |
| 150 | 57.14% | 100.00% | 57.00% | 0.7273 |
| 162 | 100.00% | 100.00% | 100.00% | 1.0000 |
| 169 | 50.00% | 75.00% | 43.00% | 0.6000 |
| Overall | 81.01% | 94.11% | 77.11% | 0.8707 |

Table 4: Performance evaluation of the system.

Once automatically identified, canonical references can have further semantic information added to them. By combining and then applying techniques of syntactic and semantic parsing to the identified references, it is possible to extract information such as the precise author name and work title, the text passage referred to, and the reference edition (either when implicitly assumed or explicitly declared).

The first important outcome of our work is that such an automatic system allows us to elicit the hidden tangle of references which links together the primary and secondary sources of a digital library. Another important outcome is that unstructured texts could be analyzed on the basis of the canonical references they contain, for example by clustering techniques. Given a consistent corpus of texts it would be possible to cluster it on the basis of the distribution of canonical references within documents in order to obtain a first topic classification.

Among the benefits of the proposed approach there is the possibility of applying it to texts pertaining to specific branches of Classics, like Papyrology or Epigraphy. Indeed in those disciplines papyri and epigraphs are also often cited by abridged references that are very similar in their structure and features to the canonical text references. In a similar way, a canonical reference parser can be trained on a particular citation style in order to tailor it to a consistent corpus of texts with consequent improvements on the overall performances.

Finally, since the task of automatic extraction of canonical references has never been explored before, we hope that in the future more resources will be available for this task (such as training datasets, golden standards, performance measure to be compared, etc.), analogous to those already existing for other more common tasks, like named entity recognition or the extraction and labeling of modern bibliographic references.

## References


Gregory Crane. 1987. From the old to the new: intergrating hypertext into traditional scholarship. In *Proceedings of the ACM conference on Hypertext*, pages 51–55, Chapel Hill, North Carolina, United States. ACM.

Gregory Crane. 2006. What do you do with a million books. *D-Lib Magazine*, 12(3).

Aron Culotta, Andrew Mccallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303, Morristown, NJ, USA. Association for Computational Linguistics.

Andreas Doms and Michael Schroeder. 2005. GoPubMed: exploring PubMed with the gene ontology. *Nucl. Acids Res.*, 33(suppl_2):783–786, July.

Andrea Ernst-Gerlach and Gregory Crane, 2008. *Identifying Quotations in Reference Works and Primary Materials*, pages 78–87.

C. Lee Giles Isaac Councill and Min-Yen Kan. 2008. Parscit: an open-source crf reference string parsing package. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Okan Kolak and Bill N. Schilit. 2008. Generating links by mining quotations. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 117–126, Pittsburgh, PA, USA. ACM.

John Lafferty, Andrew Mccallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 289, 282. Morgan Kaufmann, San Francisco, CA.

Frank Lester. 2007. Backlinks: Alternatives to the citation index for determining impact. *Journal of Electronic Publishing*, 10(2).

Andrew Kachites McCallum. 2002. MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu.

Matteo Romanello. 2007. A semantic linking system for canonical references to electronic corpora. Prague. to be next published in the proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages, held in Prague November 2007.

Matteo Romanello. 2008. A semantic linking framework to provide critical value-added services for e-journals on classics. In Susanna Mornati and Leslie Chan, editors, *ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing held in Toronto, Canada 25-27 June 2008 / Edited by: Leslie Chan and Susanna Mornati*.

David A. Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 127–136, London, UK. Springer-Verlag.

Neel Smith. 2009. Citation in classical studies. *Digital Humanities Quarterly*, 3(1).