A Semantic Linking Framework to Provide Critical Value-added Services for E-journals on Classics

Matteo Romanello[1]

[1] Ca' Foscari University of Venice
V.lo Limbraga 16, 31100 Treviso (TV) ITALY
e-mail: mailto:mromanel@dsi.unive.it

Abstract

In the field of Classical Studies the use of e-journals as effective means for scholarly research still needs to bootstrapped. This paper proposes a possible implementation of two value-added services to be provided by e-journals that is to reach this goal: reference linking and reference indexing. Both these services would contribute to make more machine-evident the hidden bond but of utmost importance which link together primary and secondary sources in this field of studies. On a technical level, this paper proposes the use of Microformats and of the Canonical Texts Service (CTS) protocol to build the semantic and non-proprietary linking framework necessary to provide scholars reading e-journals with some advanced and critical features.

Keywords: E-journals; Classics; Reference linking; Microformats ; Value-added services; Semantic Web, CTS URNs, Canonical Texts Services protocol.

1. Introduction

Without a doubt, e-journals in the field of Classical Studies are not yet so popular in this field in comparison to scholars in other scientific disciplines. It results that the use of e-journals as effective means for research purposes needs to be bootstrapped. The greater use of these electronic resources would justify greater investments by publishers in terms of both money and human resources in the production of electronic versions of printed journals or even of digital- born journals. In order to reach this aim there is need to identify what services and features really need to be provided as value-added services to scholars reading on-line publications. At the moment, in particular with respect to the Italian situation, journal publishers do not yet provide a sufficient amount of value-added services along with the electronic journals. The shifting from content holding to service providing was suggested by C. Armbruster [17]'†as a profitable model to reach an effective economical sustainability of the open access to research contents. Therefore, it is hoped that a publishers' strategy would focus on providing new and useful services for readers allowing both an increas in the use of e-journals among scholars and offering open access to a greater amount of contents.

In the field of ancient languages studies, and in particular Classical Literature and Philology, the distinction between ancient literary texts (primary sources) and modern monographies or journal articles written about them is of primary importance. Scholars' research work can be considered as lying for the most part in creating, retrieving or modifying links between these two different kinds of resources. Indeed, when they compare texts to produce new interpretations and they provide secondary sources in support of their thesis, they constantly draw new ties between primary sources, as well as between primary and secondary sources.

The first time this complex net of citations and references between primary and secondary sources tried

to emerge in a single *medium* was with the Byzantine practice of writing *scholia*. The *scholium* then made its first appearance as a new practice as a result of an emergent writing support: the *codex.* The *codex* left ancient erudite scholars more blank space all around the main writing space than the papyrus scroll in order to write comments and remarks *a latere* of the text itself (all around the text itself). This new practice led to the need for a graphical signs system to link together a given text passage and the erudite comments referred to [24]‘. Thus the *scholia* may probably also be considered an embrional kind of hypertext. At the moment, the Web gives us the technical means to get closer again, in a truly digital context, to primary and secondary sources by creating a sort of big *scholium* on a Web scale. And this will be one of the most interesting challenges for cyberscholarhip-related researches in the near future.

Therefore, the aim of this paper is to discuss both the characteristic features and the actual design of a Microformat vocabulary set to encode references to ancient literary texts within web documents in order to achieve a semantic linking system between primary and secondary sources. From now on such references will be referred to as “canonical references” to distinguish them from the citations of modern texts: they are references to discrete corpora of ancient texts that are written by scholars in a canonical citation format which often implies abridgements. The paper firstly describes the technical requirements of such a semantic linking system identified by examining the characteristics of canonical references. Then the paper shows how Microformats can be suitably imployed to encode such canonical text references thus realizing a linking system for primary and secondary sources which has to be semantic, open-ended and cross-language. T*he implementative details of the proposed Microformat will be presented and discussed, along with some examples of its use. Some examples* of value-added services that could be provided by journal publishers will be built upon the outlined linking system. In particular the paper focuses on the navigation services according to the distinction made by Armbruster between certification and navigation services, since services of this kind seem to be the most desirable for scholars of classics. Specifically these are the reference linking (where by “reference linking” I mean the capability for a user to move directly from a text reference to its source) and the reference indexing of articles published by scholarly journals.

2. Methodology: loosely vs tigthly coupled approach to realize a linking system between primary and secondary sources

Before presenting the methodological approach adopted in this paper to solve the big issue of bringing primary sources close again to secondary ones by using a semantic linking system, it is necessary to describe the actual scenario and then how the desidered scenario would look.

2.1 Scenarios

a) Jane is a scholar of classics and she is doing some reasearches on Homer’s *Iliad,* and in particular on the fourth book. During some preparatory bibliographic researches, she experiences difficulties related to information retrieval in the field of classical studies. She would like to be able to find the journal articles and the monographies written in any language where the work of her interest is dicussed. Instead, the common search engines allow Jane only to submit language-specific queries. Thus in order to retrieve *all* the publications related to the fourth book of the *Iliad* she would have to submit one query for each of the main world languages (at least German, English, French, Italian and Spanish)Jane is disappointed about the inefficiency of searching the web through the normal search engines and gives up her on-line search.

b) John is a Jane’s colleague and has just subscribed to an e-journal on Classical Studies which in addition to open access to articles also offers some experimental navigation features such as reference linking for ancient literary texts. When reading on-line articles John can read directly on the screen the original text

of each text passage of an ancient work cited in the body of the article. Morever he is able to retrieve all the available editions or translations for a given text passage, comparing critical editions or finding related resources (illustration from ancient pottery, reviews of articles or books concerned with that specific text passage etc.). Furthermore readers can browse all the journal issues using a reference index (namely an *index locorum)* which covers all the articles published by the journal itself . By using this tool John is able to find all the articles related to for example Aeschylus, the author he is focusing on, and in particular the exact lines of *Agamemnon* he is specifically investigating.

2.2 State of the art and proposal of an implementative solution

Currently, in the field of Classical Studies the secondary sources published electronically do not go beyond an incunabular stage, according to the defition of *digital incabula* given by N. Smith [27]‘. A few on-line publications are provided as HTML, whereas most of the journals just keep publishing a binary file reproducing exactly the printed issue (generally in PDF format). Nonotheless existing projects aimed at building electronic corpora of ancient languages most of the time harness an internal linking system that, on one hand permits reaching a worthy degree of hypertextuality but on the other hand, however, just produces a fairly closed system of hard-linked resources. Within electronic secondary sources the references to canonical authors and texts (that we referred to above as “primary sources”) are mostly not encoded, but in a few cases are marked up as simple links to stable on-line resources such as digital libraries and electronic corpora. The issues resulting from such a situation have been briefly summarized in the first scenario described. Abbreviations and implicit statements are not expressed in a machine-readable format and thus constantly require human disambiguation capabilities. Such references are treated by robots and information retrieval tools merely as strings without the slightest semantic markup that if present would permit a metadata-aware information processing. Indeed within secondary sources the references to ancient authors’ texts are expressed most of the time by abridgements, such as Hom. *Od*. IX 1 or 145. But these abridged notations are unmanageable when using current text retrieval systems, because, for instance, a query with a single Greek letter, that in the context of a Homeric citation is the book reference, leads to poor precision (it can mean something else, e.g. divisions in paragraphs) and recall (there are other ways to indicate the same book, e.g. by roman numbers).

```
<!-- Plut. Sol. 19.1 Canonical Text Reference from
http://www.stoa.org/projects/demos/article democracy development  -->

<a class="citation" target="_blank" href="http://www.perseus.tufts.edu/cgi-
bin/ptext?lookup=Plut.+Sol.+19.1">Plut. <em>Sol.</em> 19.1</a>
```

Figure 1: Example of hard-linked canonical reference to a passage of Plutarch’sSolon encoded following a tightly coupled approach.

To sum up, at the moment the references to primary sources contained inside electronic secondary sources are rendered as hard-links through a tightly coupled linking system or even are not encoded in a machine readable format and thus still just replicate the references contained in printed texts. Notwithstanding, there is not yet a shared standard or a well established best practice to determine how to encode references to texts stored in any given corpus, within an (X)HTML document. On top of that, digital corpora of primary sources do not use a common protocol which would guarantee interoperability among different collections of texts and would allow the creation of links between primay and secondary sources. Indeed, w*hat has to be avoided as much as possible is the use of several solutions peculiar to given projects and un-interoperable with each other. For that reason it is necessary to find both a common protocol to access collections of texts and a shared format to encode canonical references within web on-line resources.*

Therefore, the desired linking system for primary and secondary sources is required to be:

- open-ended*: it should be possible to link and retrieve other resources related to a given author or work as soon as they appear on the Web. Each link would be resolved into an open-ended, and therefore potentially infinite, number of on-line resources;*
- interoperable*: it should guarantee the reuse of data and the interoperability among web applications that use different communication protocols and interfaces;*
- semantic and language-neutral*:* such a linking system should allow to identify each author, work and edition of a work with a unique identifier rather than with a language-dependent name. If an author is univocally identified it is possible to map the name of the same author written in different languages to that unique identifier. But it is impossible to do the reverse.

In order to implement such a linking system I adopted here a loosely coupling approach. The canonical references and the primary sources referred to have been assumed to be two separate sets that should be linked together using a loosely rather than tightly coupling approach. This kind of coupling is realized by distinguishing three different layers from each other: a) the layer of metadata references contained in the web pages; b) the layer of applications or agents capable of understanding such references in order to provide new navigation and information retrieval services; c) the layer of service providers, such as collections and digital libraries containig the texts referred to by canonical texts references. Furthermore, the implementation of loosely coupled linking system such as this is made of two main components: a metadata embedding technique to mark up directly on the page the canonical references in a machine readable format, and a web protocol in order to create from the metadata embedded in the web page some dynamic links to other resources made available by different service providers.

The adopted approach was inspired by two interesting experiences that achieved a high adoption rate on the web. The first one is the OpenURL ContextObject in SPAN (CoinS) [12]‘ which allows embedding in HTML the metadata necessary to construct a link compliant with the OpenURL protocol. This solution of marking up citations and references to any publication allows some context-sensitive services of reference linking to be provided. Indeed, a user browsing a web page with a metadata-aware agent is able to switch directly from a bibliographic reference, for instance, to the corresponding record in a library OPAC or to an electronic version if available. Secondly, in the area of chemistry I found an interesting example of a loosely coupled linking system similar to that which is being proposed in this paper [29,30]‘. In electronic publications in this field a shared and non-proprietary set of identifiers to refer to chemical substances is currently being used: the IUPAC International Chemical Identifier (InChI) [20]‘. The availability of this set of identifiers permits one to build a system which links together pieces of information related to the same chemical element or substance simply by embedding in HTML the correspondent unique identifier. The solution for metadata-embedding proposed by Egon Willighagen is to use RDFa and a client-side script in order to link dynamically different on-line resources concerned with the same chemical stucture, such as blog posts or a PubMed journal article.

Now, in order to implement a similar solution in the field of Classical Studies I suggest the use of a) the *Catalog of Greek Literary Works* (and other similar catalogs), a CTS URN scheme that leverages the existing Thesaurus Linguae Graecae Canon of Greek authors [15]‘ aimed at univocally identify canonical authors, works and also different exemplars of the same work (e.g. critical editions and translations); b) an ad hoc Microformat to embed canonical metadata references in HTML elements; c) open protocols to provide some value-added services that use the semantic information embedded in microformatted references, such as the CTS protocol to retrieve texts or part of them from a distributed digital library of classical texts. This solution seems to satisfy the above identified requirements of the desired linking

system. Indeed the open-endedness is guaranteed because the links between resources are created dynamically by a metadata-aware agent on the client-side, rather than on the server-side as currently happens most of the time. Furthermore the use of a unique scheme of identifiers for authors and their works instead of language-dependent named entities will guarantee the semanticness of the suggested linking system. In the next sections both the implied technologies, Microformats and CTS protocol, will be briefly presented.

2.3 Embedded metadata layer: Microformats

Microformats are one of the most cutting-edge technologies from Web 2.0 aimed at the markup of structured data using HTML tags. They are currently developed at the developers community through a specific wiki, where it is possible to find principles and guidelines to be followed during the design process of new Microformats. Nevertheless, the development of new Microformats seems to be quite discouraged by the community itself. Types of data that are currently being encoded using Microformats include geographical data (geo), web feeds (hFeed), personal profiles (hCard), relationships (XFN), reviews (hReview), events (hCalendar), curricula vitae (hResume). Microformats' success is due particularly to their use in blogs and social networks and demonstrated how semantic data could be easily embedded inside compounds of Plain Old Semantic Html (POSH) without mixing content with presentational features, which are instead managed through Cascading Style Sheets (CSS).

2.4 Unique identification and communication layer: the CTS Protocol

The main features of projects that started in the past by building electronic *corpora* are so strictly dependent on the available technologies, that they can be thought in terms of diachronic evolution. Indeed, as pointed out by Neel Smith, "the TLG founded in 1970s is a response to the potential of mass storage and rapid retrieval in digital media" and "the Perseus project founded in 1980s is a response to richer media and higher-level data structures". Now, a distributed digital library built upon various web-oriented protocols and standards would probably be the response to the technological convergence of XML-related technologies, an easier web service communication over interfaces designed in conformity with a Representational State Transfer (REST) architectural style [18]`†and an infrastructure taking some advantages from the distributed architecture of the web itself.

The Canonical Texts Service protocol, developed among by others by Neel Smith and Chris Blackwell at Harvard's Centre for Hellenic Studies, allows the joining together different digital repositories of TEI-encoded texts, providing a common protocol to make these collections interact with each other as a single distributed library. The protocol lies in the conceptual model described by the Functional Requirementes for Bibliographic Records (FRBR) distinguishing between a work and its different exemplars, while introducing some slight differences, e.g. defining the term "workgroup" instead of "author". One of the important features of the protocol is that it allows one to reach a higher granularity when accessing documents hierarchically and supports the use of a citation scheme referring to each level of the entire document hierarchical structure. In addition, it makes it possible to distinguish different exemplars (namely instances) of the same text. *Digital libraries as well as libraries of printed texts need catalogs and authority lists to make the information they contain easily retrievable. To this end* the CTS protocol defines a scheme of Uniform Resource Names (URN), called CTS-URNs, to identify univocally authors and works contained in corpora of canonical texts that could be accessed as CTS-repositories. Such a semantic authority list could be defined for each corpus of XML texts. Therefore the protocol is in turn built upon another ancillary protocol, the Registry Services Protocol which provides "an automated interface to authority lists". The distributed and wide scope nature of Registry Service protocol allows the creation of several lists of identifiers that are organized in registries. These unique identifiers for authors, works

and other interesting facts (for instance geographical or mythological names) can be used as additional and significant entry points to texts stored in CTS-compliant repositories.

3. Results: a Microformat vocabulary set for canonical texts references

This section presents one of the main findings of this paper namely the definition of a Microformat vocabulary set to encode semantically canonical text references in HTML elements. Firstly, some use-cases references are examined in order to identify the requirements of a format in order to encode them. Secondly, the actual design of such a Microformat is discussed. Finally some examples of value-added services that *that could be built in the near future to take some benefits of such microformatted contents are supplied* .

3.1 Properties of Canonical Text References

First of all it is necessary to define the properties of references to canonical texts in order to think of a technical solution that really fits scholars' needs and usual practices. Since any reference may be written following, or not, a given citation scheme and implying some implicit information, it becomes difficult to operate a fully automated text analysis in order to identify references to *corpora* texts and extract semantic information from them. Examining a case history it is possible to distinguish at least three types of references: abridged (3,5-9), unabridged (1-2, 4, 10-11), implicit (4) on the basis of their appearance. Three genres could be distinguished on the basis of its meaning: references to a canonical author, to a canonical work and to a precise text passage. In addition, since references are language-dependent this fact exponentially increases the number of possible equivalent references to authors and works (Homer, Omero, Homero, Homerus and Homère are all valid names the same author might be referred to by). Furthermore, a common practice of scholars in classics is to refer to different edition shortening the editor's name (7,8) or referring to the current edition of a text just by omitting the editor's name indication. Indeed, a common practice is to leave out the editor's name indication when a work is known entirely and to indicate it instead when it is known by fragments. In particular this case makes references even more context-dependent since,when another edition becomes the most reference edition for scholars, the meaning of the reference itself changes. Otherwise, subsequent anaphoric references to different *loci* of the same texts are usually written specifying for the first time the text referred to and then indicating just the line or the number of the chapter pointed to by the author (5,6).

(1) [...] the *Politics* of Aristotle
(2) Homer and the Papyri
(3) Ar. 'Arch.' 803
(4) Protean Forms and Disguise in 'Odyssey' 4
(5) Aesch. 'Sept.' 565-67, 628-30
(6) Aesch. 'Pers.' 265, 711, 1008
(7) Hes. fr. 321 M.-W.
(8) Callimaco, 'ep.' 28 Pf., 5-6
(9) Paus. 7.23.1-3
(10) [...] sulla parodo dell'"Ifigenia in Aulide" di Euripide (ov

Figure 2: Case history of canonical references written in different languages and following several citation styles.

In conclusion, from the analysed case history results that the indication respectively, of the author and of the work referred to, should be considered at the same time as the zero degree and the minimum properties

of a canonical text reference. On the contrary, the indication of text range and editor's name appear to be additional information that contributes to setting up a precise text reference, concerned with a specific exemplar (both digital or printed) of a canonical work, such as a translation or a critical edition. Futhermore, since even the title of a work or the name of an ancient author can be considered with good reason as a reference, the Microformat to encode such references is should be able to cover as much as possible the entire case history.

3.2 Design of a Microformat vocabulary set to encode canonical text references

After the properties of canonical text references have been singled out, it is necessary to design a Microformat vocabulary set in accordance with the above specified requirements choosing the most fitting among POSH tags. The design process of new encoding formats should conform to Microformats design principles and patterns, such as embeddability and modularity, in the way desired by the developers community [16]‘.

The first Microformats design principle states that a "microformat must solve a specific problem": the problem we started from is how to link primary and secondary sources within a Digital distributed Library in a way open-ended, semantic and cross-language way. Furthermore, according to the fourth principle claiming that designing Microformats means "paving the cowpath", the proposed microformat does not change the current behaviour of the users for whom it is especially designed (i.e. scholars). Hence, our solution should fits with the actual habit of citing canonical texts outlined above since each Microformat is suitable to encode abridged, unabridged and implicit references as well.

The HTML tags used to compose the microformat vocabulary set are the most semantic among POSH elements. Since the entire canonical reference could be defined in a loose sense as a citation, a <cite> element was used as container element. As highlighted by the chunk of code reproduced in Fig. 3, the value of class attribute of each element indicates the property expressed by the element itself as implied by the Microformats class-design-pattern [5]‘. I propose three Microformats here that may be combined to encode almost the entire range of possible references individuated above (Fig. 2). In particular, the decision to split the vocabulary set into three different Microformats on one hand responds to needs of embeddability and modularity, and on the other hand corresponds to the earlier adopted classification of references by meaning. The "ctauthor" (Fig. 3, line 3) is an elemental Microformat and should encode references just to a canonical author. The "ctwork" (line 6) belongs to the same kind and is suitable for references to canonical works where the author's name is implicitly contained. The author statement is made machine-readable thanks to the structure itself of the CTS URN correspondent to a work which contains also the CTS URN of its author (e.g. "urn:cts:greekLit:tlg0012.tlg001" refers to Iliad and and the same time contains the CTS URN of Homer, "urn:cts:greekLit:tlg0012"). Both "ctauthor" and "ctwork" Microformats was designed according with the abbr-design-pattern [1]‘† which suggests to use an <abbr> element to encode data that are provided in both machine-readable and human-readable format. Indeed the title attribute of <abbr> element is used to supply the same data contained in the element in a machine-readable (and -parsable) format, while the inner . Finally, "ctref" (lines 3-10) is a compound Microformat to encode a complete canonical reference. Therefore it requires as mandatory properties both a "ctauthor" and a "ctwork" element and a "range" property which supply a numerical range of the text sections referred to. In addition it allows to specify an edition statement property (line 9) which is designed following the abbr-name pattern and encodes the editor's name either in a complete or abridged form.

To sum up, the proposed Microformat allows content providers to combine this encoding format with an internal linking system if already present (Fig. 3). The microformatted references embed the metadata necessary to make them machine-readable and at the same time this semantic encoding is reached regardless

of the external appearance of canonical references which is managed through CSS. Therefore, scholars are not requested to change their citation practices because the suggested Microformat vocabulary set is able to cover any citation format and content providers could even integrate it easily into their project-specific linking framework. Ultimately, the use of such a Microformat does not exclude the possibility of hard-linking the same references to resources on the web, it just provides the semantic information necessary to also build on top of them a loosely coupled linking system.

```
1  <a class="citation" target="_blank" href="http://www.perseus.tufts.edu/cgi-
2   bin/ptext?lookup=Plut.+Sol.+19.1">
3   <cite style="" class="ctref">
4     <abbr class="ctauthor" title="urn:cts:greekLit:tlg0007">Plut.</abbr>
5      <em>
6       <abbr class="ctwork" title="urn:cts:greekLit:tlg0007.tlg007">Sol.</abbr>
7      </em>
8      <abbr class="range" title="19.1">19.1</abbr>
9      <abbr class="edition" title="Bernadotte Perin"/>
10   </cite>
11 </a>
```

Figure 3: Example of the hard-linked canonical reference of Fig. 1 after it has been reencoded using the proposed Microformat vocabulary set.

3.3 Benefits of microformatted canonical text references

Various types of resources often containing canonical references that could be microformatted in addition to both monographies and e-journals articles, include books reviews, conference announcements, descriptions of heterogeneous resources (such as illustrations on pottery), metadata of bibliographic records. All those resources could be suitably aggregated if the contained references to primary sources could be properly encoded.

Once some Microformatted contents are available on the Web, an engine for targeted search, modeled on existing services that allow one to search among information tagged using a given microformat (e.g. Technorati's Kitchen supporting in particular the search among hCard, hCalendar and hReview microformats), will make possible it to retrieve resources pertinent to a given author or work with greater precision and recall, giving the scholars information retrieval tools infinitely more precise than traditional web search engines. The increased precision is granted by the capability of CTS URNs of addressing even a single part/section (i.e. book, chapter, paragraph, line, verse ecc.) of an XML-encoded literary text. Furthermore microformatted contents are searchable even through a not-semantic search engine (like Google) because of their bidirectionality [25]'. This property is due to the fact that inside a microformatted canonical reference some URNs are directly embedded in HTML and thus a textual search engine will also retrieve them.

In the near future useful desktop applications allowing the reuse of microformatted data extracted from web pages may be created. It is likely, for instance, that an application could allow scholars to manage collections of the most frequently used references and to export them formatted according to a given citation style (or in a localized or abridged, rather than unabridged, form) to a wordprocessor. One interesting example of applications like this which work with citations and bibliographic references of modern publications is Zotero. Zotero allows one to extract from web pages and then manage or export in different formats the bibliographic references expressed using CoinS, the above mentioned technique of metadata embedding which is very like a Microformat even if it is not properly a Microformat proper because it was developed outside that community and without following the Microformats patterns and design principles.

But for scholars on classical literature the most useful interface functionality to provide, and the only one capable of giving readers of the Digital Library a richer and more meaningful experience, is without a doubt the reference linking. As soon as a significant number of digital libraries exposing a CTS-compliant

interface are available over the web, it will be possible to build client-side applications giving readers the ability to move directly from a text reference to its source, and to read or compare different critical editions of the same texts. The prototype of an e-journal providing such a service will be described in the next section

3.2 E-scholia: prototype of an e-journal on classics with a feature of reference linking for primary sources

E-scholia is the prototype of an e-journal on classical philology and was conceived to realize the scenario described in section 2.1.b. Its aim is to implement a reference feature to be provided as a critical value-added service to scholars who read on-line journal articles. The scholarly articles published by E-scholia are XML-encoded according to the TEI P5 specification. This makes it possible to both display the same article in several output formats (plain HTML, HTML with rich user interface and PDF) and to provide the users with some interesting navigation features made possible by the text encoding. The canonical references contained in each article and previously marked up with TEI-XML, are then encoded with the proposed Microformat vocabulary set. Thus each reference is mapped dynamically onto CTS-compliant requests that will receive back from a CTS repository - providing that someone would be available – an XML-encoded response which contains the requested text passage. This way each reference to a primary source becomes a direct link to the text referred to. Fig. 4 (n. 2) shows the text corresponding to the canonical reference Hom. *Od*. 2.272 to be dynamically retrieved by querying a CTS repository available on the web and then displayed to the user in a small text box. A client-side script constructs on the fly the CTS query using the metadata extracted from the microformatted reference. Once on the web some CTS repositories are available providing other critical editions and translations of the Homer' Odyssey the reader using such a reference linking feature will be able to browse and read them starting with the reference Hom. *Od*. 2.272. Actually the stability and persistency of existing digital libraries sharing their raw data as a canonical texts service needs to be considerably improved. Once this is improved it will be possible to provide such advanced features as stable valued-added services for e-journals on Classical Studies.

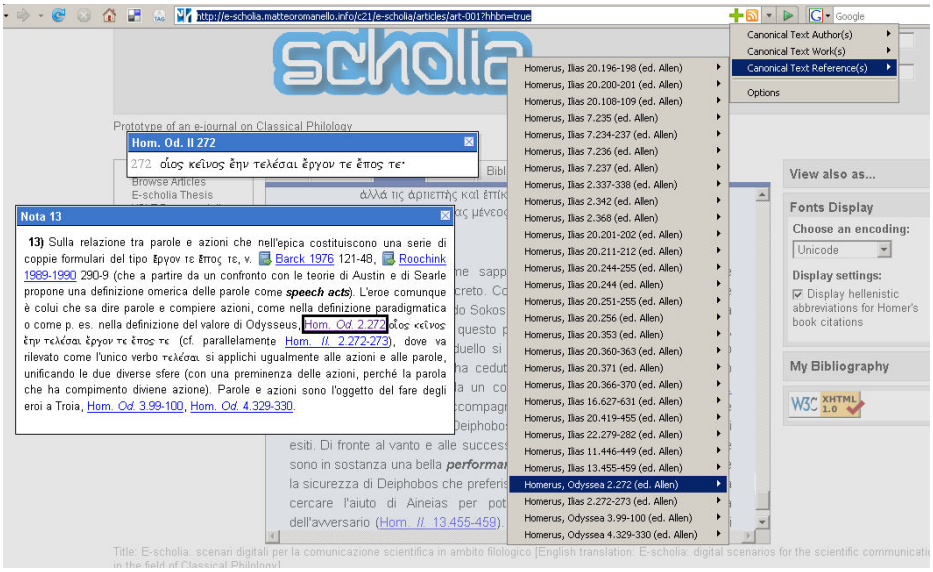

Figure 4: Screenshot of E-scholia prototype of an advanced e-journal on classical philology. Canonical texts references (1) becomes dynamically links to the full-text of the passage referred to which is retrieved from some CTS repository available on the Web. Furthermore an extended version of Operator extension for Mozilla Firefox allows to detect the meaning of microformatted references and to perform upon them some actions on the client-side (3).

The main benefit of using a Microformat instead of a non-standard encoding format to markup canonical references is that the semantic metadata contained in the reference allow the implementation of a server-side feature which does not exclude the possibility that the reader can use other client-side agents to parse the same microformatted references and to find related information. To demonstrate how this could be accomplished I extended Operator [21]‘, a Javascript extension written by Michael Kaply for the popular open-source browser Mozilla Firefox, that will be fully integrated within version 3.0 of the web browser. Operator works by parsing the entire page from which it extracts the microformatted contents and then suggests the available actions to the user (Fig. 4, n. 3). I extended Operator adding just a few lines of code in order to add support for the proposed Microformats vocabulary set. The result is a browser aware of Microformatted references and able to act upon them. The only action supported at the moment is to searching both the web and among del.icio.us’ bookmarks to find a given CTS URN: this is accomplished by appending dynamically a CTS URN as query parameter to the standard search query URL of those services (e.g. CiteULike’s query URL to search among tags is “http://www.citeulike.org/search/all?q=tag:<query parameter>”). Given that the support for this action was added just by virtue of example, several services could be built on the basis of CTS URNs, among them there some reference indexing services that are outlined in the next section.

3.3 Outline of a reference indexing service for e-journals on classics

A service of semantic reference indexing is another example of services that could be provided in order to benefit from such microformatted references. Indeed the support for such a service can be added to already existing client-side agents that are aware of Microformats. This service is supposed to take as input a CTS URN and to return a list of resources related to the supplied URN regardless of the language they are written in. In particular a reference indexing for e-journals will return all the indexed articles that contain in the title or in the text body one or more references to a given author, work or textual passage. The glue between the layer of such a reference indexing service and the layer of metadata contained in canonical references is granted by a client-side application as for instance a browser add-on (see Fig. 5).

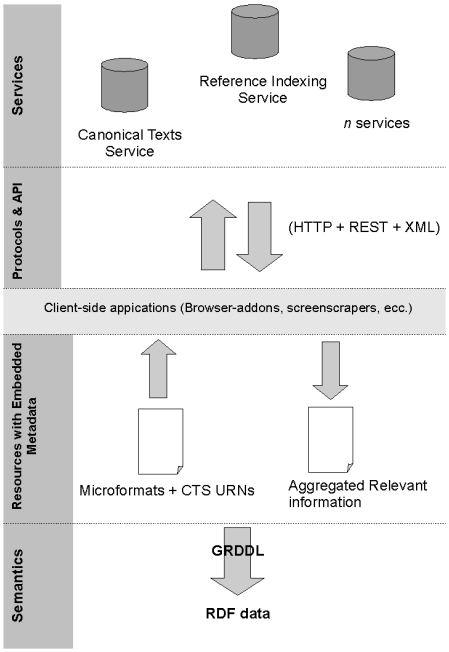

Figure 5: Separation of logical layers in the proposed linking framework between primary and secondary sources.

Currently, scholars of classics often use the electronic version of L'Année Philologique (Aph) [10]‘†for their bibliographic researches which allows one to search at one time among all the publications summarized within the printed version of the well-known bibliographical index. Publications are indexed by the ancient authors they focus on (if specified) using a latin thesaurus wich partially avoids the above described drawbacks due to the fact that named entities are language-dependent (see scenario 2.1.a). Users can search the index using the Latin name of an ancient author as the search key. Furthermore the information retrieval system of APh does not allow users to search for publications related to a specific text passage. Instead, a semantic reference indexing service should allow one to search, for instance, for all the articles concerned with the first line of Homer's Iliad and Odyssey since these articles are mapped to the CTS URNs "urn:cts:greekLit:tlg0012.tlg001" and "urn:cts:greekLit:tlg0012.tlg002". Lastly, since it favours the traditional printed version of summarized articles APh does not provide the URL of articles that are also available electronically.

At the moment there are at least a top-down and a bottom-up approach. The top-down approach is to use the Reference Index protocol, a web protocol which is being developed at CHS and could be profitably employed to create a semantic reference indexing service for on-line published articles. This protocol actually allows the creation of a list of entries where each entry corresponds to a resource and is associated with a unique identifier which could be a CTS URN or a canonical reference. Since such a service might take as an input a CTS URN and should return a list of related journal articles, it will be fully interoperable with the suggested Microformat. Although it is undoubtably the most powerful solution, this approach presumes that the Reference Index protocol is widely adopted by content providers such as e-journals.

Instead a bottom-up solution to provide such a reference indexing service for published e-journals articles would properly exploit the tagging feature provided by the most part by on line platforms for the social bookmarking. The basic assumption is that tags could also be semantic. In order to tag semantically an article which focuses on the Homer's *Iliad* it is just enough to use it along with tags as "Homer" and "Iliad" the CTS URN correspondent to the given work (i.e. urn:cts:greekLit:tlg0012:tlg001, for the *Iliad*). This application of the semantic tagging technique was inspired by some experiments recently done by Sebastian Heath [26]`†and Gabriel Weaver [28]`†in the same research perspective. Among the experiments I began conducting there is the semantic tagging on CiteULike of some bibliographic citations from e-journals on classics by using CTS URNs as tags instead of just named entities to describe the topic of the paper declared in the title [4]`. Once some available bibliographic records are properly tagged, it is possible to figure out, for example, an application (even acting as a web service) which: a) harvests several CTS repositories containing the XML encoding of primary sources; b) displays the original text to the reader; c) with the respect to the actual reading context retrieves from a service such CiteULike some bibliographic entries concerned in particular with the text passage that is being read which have been previously semantically tagged; d) displays the relevant aggregated information to the user in a totally transparent way: thus he is not required to know what CTS URNs are.

Among the advantages of a bottom-up approach such as the semantic tagging we need to mention the fact that it does not imply any modification to be made directly to existing contents, such as lists of bibliographic records, but it just simply implies the reuse of raw contents that may be created in a collaborative way as through a social application from the Web 2.0.

4. Discussion: Microformats and CTS URNs suitability

With respect to the adoption of Microformats to implement a semantic linking framework it is noteworthy that their suitability is confirmed by the Rule of Least Power: this is the reason why they are preferred among the others to RDF [14]‘. Indeed the data complexity of canonical references is not enough to make necessary at all the capability of RDFa [13]‘†or eRDF [7]‘†to embed within elements and attributes of

HTML such complex semantic data that have to be expressed by using multiple RDF vocabularies. Due to their rapid and wide success Microformats undoubtably contributed to make the topic of reaching the Semantic Web more popular and to draw attention to it again. Indeed someone [22]‘†seems to have appropriately coined the term “lower-case semantic web” or also “real world semantics” to describe them, since at the moment are considered a sort of bottom-up way to the Semantic Web itself. The use of Microformats to express semantic data may be considered as a starting point toward the upcoming adoption of more powerful and semantic web-oriented technologies, such as RDF. Moreover, the forward-compatibility with a more powerful format such as RDF can be reached through the use of some already existent vocabularies and ontologies, such as *Expression of Core FRBR Concepts in RDF* [8]‘*, Citation Oriented Bibliographic Vocabulary* [3]‘†and *Dublin Core Metadata Element Set* [6]‘. Thus microformatted references can be extracted as RDF triples by using these vocabularies to express semantic data, after have been processed by a GRDDL transformation [19]‘.

Another fact which confirms Microformats’ suitability is that both of the next versions of the two most popular web browsers (i.e. Internet Explorer 8 and Mozilla Firefox 3, arrived yet at its Beta 5 version) have been announced as featuring support for this technology contributing thus to redefining even the role played by web browsers. Indeed web browsers are actually turning into platforms for content aggregation rather than mere tools to display static pages. There are some interesting examples, such as the social web browser Flock [9]‘, those browsers are conceived on the basis of the needs of the community of users they are expected to be used by. Scholars in classical studies could in the near future take benefit from a browser customized to fit better their own needs and providing integration for some useful tools.

Furthermore, with respect to the development policy of new microformats it is noteworthy that although it is fully controlled in quite a centralized way by the Microformats community, contrasting in some way with the Web’s decentralised nature itself, it undoubtably prevents a large amount of redundancy encouraging discussion and agreement on new standards proposals. Currently, the Microformats developers community is working on hBib a format to encode references to modern publications that was previously called secondary sources according with a common distinction in the field of classical studies. Therefore it is time to start a discussion on the Microformats Wiki [11]‘†aimed at defining a common format to encode references to primary sources, and in particular canonical references.

Finally, it could be asked “and why did you suggest building such a linking system for primary and secondary sources upon the CTS protocol and CTS URNs, although they are not yet widely adopted?”. The first reason is that currently the CTS is an effective and unparalleled protocol in the field of classical studies. Furthermore it fits perfectly the need of scholars of classics referring and accessing hierarchical sections of canonical works within digital libraries through unique identifiers [23]`. Secondly, it is built entirely upon common and fully interoperable standards and responds to the Web’s decentralized nature. In fact, recently the Perseus Project started implementing a CTS-compliant web interface for its digital library which is probably the widest free accessible digital collection of classical texts. The Perseus’ choice of allowing potentially thousands of web applications to use its raw XML data to provide new services is destined to increase its popularity rather than reduce it and seems to bode well for a wide-scale adoption of the CTS protocol itself. After that the MultitextHomer will be completed [2]`, one of the first big projects to be built, among other technologies, just upon the CTS protocol, it is hoped to become a model for other “monographic” projects aimed at digitazing the whole history of the text of the works of a single ancient author. This is a great opportunity for cyberscholarhip to realize a distributed and freely accessible digital library from which it will be possible to start improving both the quality of digital critical editions and their effective usefulness for scholars.

5. Conclusions

In conclusion, this paper tries to show what benefits that could be gained by using a linking framework between primary and secondary sources based on a loosely coupled approach, and implemented by using Microformats combined with CTS URNs. In particular, the benefits addressed in this paper, would increase the efficiency of information retrieval tools and enhance the number of services it is possible to provide compared to using a tightly coupled linking system. When some effective value-added ~~serivec~~ services such as the reference linking or the reference indexing will be provided by e-journals, scholars themselves are hoped to be more encouraged to use the electronic publications for their discipline-specific purposes. However, some improvements are expected in particular with regard to the automatic creation of microformatted references, since a model just based on a fully manual markup will not scale. To this end some Natural Language Processing techniques could be profitably exploited in the near future, such as finite-state automata, named entities disambiguation and semantic analysis of the writing context where such references appear.

6. Notes and References

[1] abbr-design-pattern - Microformats. [cited 12 May 2008]. Available from world wide web: <http://microformats.org/wiki/abbr-design-pattern>.

[2] Center for Hellenic Studies -The Homer Multitext Project. [cited 10 May 2008]. Available from world wide web: <http://zeus.chsdc.org/chs/homer_multitext>.

[3] Citation Oriented Bibliographic Vocabulary. [cited 10 May 2008]. Available from world wide web: <http://vocab.org/biblio/schema>.

[4] CiteULike: mromanello56k's library. [cited 10 May 2008]. Available from world wide web: <http://www.citeulike.org/user/mromanello56k>.

[5] class-design-pattern - Microformats. [cited 9 May 2008]. Available from world wide web: <http://microformats.org/wiki/class-design-pattern>.

[6] Dublin Core Metadata Element Set, Version 1.1. [cited 10 May 2008]. Available from world wide web: <http://dublincore.org/documents/dces/>.

[7] ERDF - GetSemantic. [cited 10 May 2008]. Available from world wide web: <http://getsemantic.com/wiki/ERDF>.

[8] Expression of Core FRBR Concepts in RDF. [cited 10 May 2008]. Available from world wide web: <http://vocab.org/frbr/core>.

[9] Flock Browser - The Social Web Browser. [cited 12 May 2008]. Available from world wide web: <http://flock.com/>.

[10] L'Année philologique. [cited 12 May 2008]. Available from world wide web: <http://www.annee-philologique.com/aph/>.

[11] Microformats Wiki. [cited 10 May 2008]. Available from world wide web: <http://microformats.org/wiki/Main_Page>.

[12] OpenURL ContextObject in SPAN (COinS). [cited 10 May 2008]. Available from world wide web: <http://ocoins.info/>.

[13] RDFa Syntax: A collection of attributes for layering RDF on XML languages. [cited 10 May 2008]. Available from world wide web: <http://www.w3.org/2006/07/SWD/RDFa/syntax/>.

[14] Resource Description Framework (RDF) / W3C Semantic Web Activity. [cited 10 May 2008]. Available from world wide web: <http://www.w3.org/RDF/>.

[15] Thesaurus Linguae Graecae - TLG. [cited 13 May 2008]. Available from world wide web: <http://www.tlg.uci.edu/>.

[16] Allsopp, John. *Microformats: empowering your markup for Web 2.0*. Friends of ED, 2007.

[17] Armbruster, Chris. Society Publishing, the Internet and Open Access: Shifting Mission-Orientation from Content Holding to Certification and Navigation Services? *Social Science Research Network* 2007. Available from world wide web: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=997819>.

[18] Fielding, Roy T., and Richard N. Taylor. Principled design of the modern Web architecture. *ACM Trans. Inter. Tech.* 2, 115-150.

[19] Hausenblas, M., W. Slany, and D. Ayers. A Performance and Scalability Metric for Virtual RDF Graphs. In *3 rdWorkshop on Scripting for the Semantic Web (SFSW07), Innsbruck, Austria*, 2007.

[20] International Union of Pure and Applied Chemistry. The IUPAC International Chemical Identifier (InChITM) . [cited 2 May 2008]. Available from world wide web: <http://old.iupac.org/inchi/>.

[21] Kaply, Michael. Operator. *Mike's Musigns*. [cited 12 May 2008]. Available from world wide web: <http://www.kaply.com/weblog/operator/>.

[22] Khare, Rohit, and Tantek Çelik. Microformats: a pragmatic path to the semantic web. In *Proceedings of the 15th international conference on World Wide Web*, 865-866, [Edinburgh, Scotland]: ACM, 2006.

[23] Mimno, David, Alison Jones, and Gregory Crane. Hierarchical Catalog Records: Implementing a FRBR Catalog. *D-Lib Magazine* 11, October 2005. [cited 13 May 2008]. Available from world wide web: <http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/october05/crane/10crane.html>.

[24] Reynolds, Leighton D. *Scribes and scholars : a guide to the transmission of Greek and Latin literature*. 2nd ed., rev. and enl. [Oxford]: Clarendon Press, 1974.

[25] Sebastian, Heath. A DNID that can find itself. *Domain Name Identifiers* November 2007. [cited 10 May 2008]. Available from world wide web: <http://dnid-community.blogspot.com/2007/11/dnid-that-can-find-itself.html>.

[26] Sebastian, Heath. Wikipedia and Google like DNIDs. *Domain Name Identifiers*. [cited 10 May 2008]. Available from world wide web: <http://dnid-community.blogspot.com/2007/11/wikipedia-and-google-like-dnids.html>.

[27] Smith, Neel. TextServer: Toward a Protocol for Describing Libraries. *Classics@* 2, 2004. Available from world wide web: <http://www.chs.harvard.edu/publications.sec/publications.sec/classics.ssp/classics_2_smith.pg>.

[28] Weaver, Gabriel. Adding Value to Open Scholarly Content. *Gabriel Weaver's Personal Blog* March 2007. Available from world wide web: <http://blog.gabrielweaver.com/2007/03/test-post.html>.

[29] Willighagen, Egon . Chemical RDFa with Operator in the Firefox toolbar. *chem-bla-ics* June 2007. [cited 2 May 2008]. Available from world wide web: <http://chem-bla-ics.blogspot.com/2007/06/chemical-rdfa-with-operator-in-firefox.html>.

[30] Willighagen, Egon . SMILES, CAS and InChI in blogs: Greasemonkey. *chem-bla-ics* December 2006. [cited 2 May 2008]. Available from world wide web: <http://chem-bla-ics.blogspot.com/2006/12/smiles-cas-and-inchi-in-blogs.html>.