New Value-Added Services for Classics E-Journals

Matteo Romanello, <matteo.romanello@kcl.ac.uk>

PhD Candidate in Digital Humanities, King's College London

Abstract

This paper examines issues related to electronic journal publishing in the field of Classics with a specific focus on the discussion of new value-added services for e-journals. Preliminary experimental data from a survey that explored the diffusion of digital technologies into the publishing workflow of Italian Classics journals identified two new value-added services: reference linking to primary sources and semantic indexing. This paper also emphasizes the importance of supporting citation persistence for electronic resources. Finally, it will describe the significance and overall functioning of these services and then conclude with an outline of the characteristics of the main technical components needed for an e-journal implementation that provides these identified services through an extension of the Open Journal Systems (OJS) platform.

1 Introduction

E-journals are far less widespread in the Humanities than in the Scientific, Technical and Medical (STM) field, not only because of a lag in Humanities journals' development but also because electronic publishing technologies fit very differently within various disciplines. The benefits of e-publishing for traditional journals are significant, including superior knowledge dissemination and the ability to reach a far wider audience. Therefore this paper aims at designing a new model for Classics e-journals that is specifically tailored to Classicists' needs.

The first two sections describe and examine survey results that assessed the degree to which digital technologies have entered the production workflow of traditional Classics journals. This survey was carried out with the support of the “Rinascimento Digitale” foundation and involved direct participation of several editors of Italian Classics journals. Although the survey only examined Italian journals, the results provided insight into tendencies of the whole field and were also consistent with related studies in other countries. It also allowed us to explore the current situation of Classics journals and thus to identify both critical issues and opportunities to enhance the overall quality and value of serial electronic publications in the Humanities.

The third section examines the set of value-added services identified by the survey that could possibly be used to promote the e-journal as a valuable medium for scholarly communication in this field. Reference linking (section 4.2), semantic indexing (section 4.3) and citation persistence (section 4.4) were all found to be services that could be provided by e-journal publishers that would also likely be considered as useful and desirable by end users (i.e. scholars).

Finally, the last section analyses the main technical components needed for a real world implementation of the e-journal model previously described in terms of its theoretical implications. In particular, we address critical issues at each stage of production, including journal workflow management in a digital environment, the automatic production of meta-
data, and the exposure of contents on the Web.

2 Classics e-journals: the state of the art in Italy

2.1 Survey Goals

The survey’s main purpose was to determine what was considered state of the art for e-journals in Italy. Given the low number of journals, however, that had both an electronic version and editors that participated in the survey, the collected data about current e-journal features cannot be considered representative. Nonetheless the survey is still illustrative since it illuminates some of the obstacles to providing electronic access to journals in Classics and also shows to what extent electronic tools are being used in editorial workflow.

2.2 Methodology

As a preliminary step, we selected an analytic sample of journals for our survey by using the Année Philologique (APh), the international reference bibliographic index for Classics and Philology. The Centro Italiano dell’Année Philologique (CIAPh) serves as the national editorial board for this index and annually reviews 276 Italian serial publications for its contribution to the APh's. The serials reviewed by the CIAPh, however, also include academic bulletins that cannot strictly be considered journals. The initial list from the CIAPh was thus filtered by using a qualitative index, the European Research Index for Humanities (ERIH), in order to select only those titles that are considered most valuable by scholars themselves. ERIH provides a category classification (i.e. “A”, “B” and “C”) for a catalogue of journals that depends on the journal’s compliance with a set of qualitative standards established by an international expert panel. The final result was a significant analytic sample of 55 Italian journals that were also included in the ERIH index.

The main areas of interest covered by the survey were:

1) the features of the electronic version of the journal (when available);
2) the journal's copyright policy;
3) the use of electronic tools to support the editorial workflow;
4) willingness and obstacles to innovation.

The questionnaire contained 31 items: three were open questions (covering in particular the fourth area) and the remaining ones were multiple-choice. It was then administered electronically to the journal editors in the sample during the month of January 2008. One noteworthy detail is that at that time only 26 of the 55 journals were present on the Web either with an electronic version or an institutional web page.

2.3 Results

Only 18 out of the 55 invited journal editors participated in the survey and completed the questionnaire. The log analysis revealed that in 18 cases the invitation was ignored, in 8 cases the participant opened but did not complete the process, and in 11 cases the questionnaire was only partially completed.

A first result from the survey seems somewhat contradictory. Of the journals that responded to the survey, only 3 of the 18 were available electronically, despite the fact that all 18 journals supported electronic submission of paper manuscripts and used commercial software to produce the printed journal version. This demonstrates that despite the fact
that the production workflow was already partly digital the preferred (and often the only) medium for publication still remained print.

The answers involving copyright policy revealed that only half of the journals allowed their authors to publish paper pre-prints on the Web and only 7 allowed the publication of a paper post-print. With regard to the publishing workflow, a document containing author guidelines was accessible electronically in just 5 cases out of 18, even though in all cases such a document was stated to exist. Finally, it was discovered that Microsoft Word was accepted by all of the journals as a valid format for paper manuscript submission and that the most commonly used Greek font for manuscript submission was SuperGreek (accepted by 14 out 18 journals). This is surprising due to the fact this font is not Unicode-encoded, which would help solve a number of visualization and interoperability issues. Unicode-encoded fonts were accepted only in 4 cases.

3 Understanding the Classics Field

This section will analyse the needs of the Classics field that emerged from the survey, with a particular concentration on the answers to the open questions according to the following polarities: 1) tradition/innovation; 2) precision/rapidity; 3) wide access/long term access; 4) open access/economic sustainability.

3.1 Tradition/Innovation

A first kind of dialectic between tradition and innovation can be found on the technical side. Indeed, in those rare cases where an electronic journal is available, the electronic version still imitates the printed one, constituting a so-called “digital incunabulum”.

This tendency is also reflected by the software in use and by the data formats adopted both for submission and publishing. Given that even Microsoft Word supports the encoding of additional information about document semantics or logical structure (for example by providing a template document that shows a consistent use of formatting styles that can then be properly transformed into mark-up), this happened in just one case. On the other hand, the PDF format – which substantially replicates behaviours and paradigms of the printed page – was found to be the most frequently used format for electronic publishing. It is clear now, however, that the opportunities of shifting to a fully electronic publishing environment could allow publishers to maximize the benefits of a digital workflow by producing both a print and an electronic version of journal contents.

Analysis of responses to the three open questions also illustrated that editors considered many Classicists limited use of digital technologies to be a factor that discouraged further investment into the realization of e-journals. Moreover, participants to the survey numbered the following as the main difficulties of going digital: the weight of the editorial tradition, the problem of converting old journal issues to an electronic format, and the difficulty of establishing e-journals as authoritative, particularly those journals that do not have a solid tradition in print publishing.

3.2 Precision/Rapidity

The promptness of the publication process is not as urgent a need for humanists as it is for those in other disciplines. Indeed, the higher obsolescence time (information ageing time) in the Humanities has at least two consequences: far less need for rapidity and a much greater degree of concern about the long-term preservation of electronic resources published on the Web. Immediacy is often simply regarded as a consequence of electronic
publishing and not as one of its most desirable benefits. Furthermore, when editors were asked: “Do you believe that the rapidity of publication allowed by digital technologies negatively affects the overall accuracy of publications?”, 5 out of 18 participants answered affirmatively, reflecting a die-hard and systematic prejudice that exists among many Classics scholars regarding the benefits of new technologies.

Rather than speed two other properties considered crucial in terms of electronic resources were long-term persistence and citation precision (or the capability of creating precise links both to and within resources, also known as link granularity). An electronic resource that does not allow users to precisely refer to its contents is considered useless in this field. Indeed, many Classicists consider the precision of bibliographic and canonical references to be an important indicator of the overall quality and scientific value of a publication. This is probably the main reason why a paginated data format (i.e. PDF) is still largely used to the detriment of a more expressive but un-paginated one (i.e. HTML), since the former allows one to use page numbering as citation schema. As a result, the aforementioned APh, the most reliable bibliographic index in this field, does not provide links to electronic versions of printed papers even when they are available, not even in its online version.

3.3 Broad access/Long term access

Electronic resources were still largely perceived as “for experts only.” As long as such resources are not reviewed (or at least referred to) by the reference bibliographic indexes (such as APh), it will be very difficult to promote the production and use of e-journals and thus scholars' awareness of them in the field of Classics.

Along with the issue of broader access there is the related challenge of digital preservation, or how to effectively provide long-term access to electronic resources. A resource that is widely accessible but extremely unreliable is not very useful. Since information ageing in the Humanities is typically higher than in other disciplines, the need to guarantee the accessibility of electronic resources over decades has become quite urgent. This issue of long-term access also highlights the need for citations to resources to be precise and stable over time as well. Given the current lack of guarantees for the long-term persistence of cited electronic resources, the current citation practice is still (and inevitably) predominantly paper-oriented.

3.4 Open access/economic sustainability

Several editors also mentioned open issues with publishers as an obstacle to going digital. These issues were mostly related to copyright, such as existing policies that do not allow electronic versions to be published. In some cases editors reported problems involving accessibility of online published contents: accessibility was defined both as the long term persistence of e-journals and the electronic resources cited within them and as open access to published research results. While the first meaning of accessibility refers to the well-known problem of “broken links”, the second one relates to the problem of finding a sustainable business/economic model to support an open access policy. From the editors' perspective, open access was seen as a highly valuable goal worth pursuing but one that also seemed to be hardly sustainable from an economic point of view. This perception fully reflects the ongoing debate and research about Open Access in the Humanities, an issue that cannot be fully addressed in this paper. It also illustrates how Humanities fields perhaps may have a hidden natural tendency to Gold Open Access but are still not very far on the road towards achieving it.
Many of the Italian editors also feared that the electronic version of a journal would cause economic damage by decreasing subscription revenues from printed serials. Several case studies\textsuperscript{3-5,11} have recently demonstrated, however, that moving to an e-journal publishing model reduces production costs by rationalizing and automating many editorial procedures. Moreover, the shift to an electronic publishing model represents a chance for journal publishers to become service providers rather than just content providers\textsuperscript{12,13}. Once journal contents are made openly accessible, subscriptions can be offered to institutions or individuals that wish to access advanced value-added services, such as those suggested in this paper.

4 Services that Matter

This survey clearly illustrated that Classics e-journals are not yet widespread and thus their use and growth needs to be promoted. Even though some advanced projects exist\textsuperscript{3,6}, they are clearly in the minority. Use and awareness of e-journals, open archives and open access in general is still far less widespread in the Humanities than in STM disciplines, a fact that can be observed simply by looking at the practice of citing electronic resources in these fields\textsuperscript{9}.

One approach to this problem perhaps might be to try and find killer applications for the Humanities,\textsuperscript{14} rather than undertaking a massive shift of the entire discipline, such as happened for Physics with the SCOAP3 initiative\textsuperscript{15}. A key question to be answered is: “What services that can be provided by e-journals will really matter for classicists?”

Our starting assumption in facing this problem is that a great part of the overall value (and benefit) of academic e-journals comes from providing services that are considered valuable by their users (i.e. mainly scholars). Indeed, the main difference between print and electronic journals lies in the layer of services provided to the user. Providing effective services could allow e-journals to move out of the digital incunabula stage, and might also foster a business model that truly makes open access economically sustainable.

4.1 Importance of Citations

This research focused on the identification of new value-added services, and in particular on navigation services, starting from the specific needs and critical requirements of Classicists. Citations play a role of primary importance in the Humanities and particularly in Classics, a point that has been previously made by Gregory Crane\textsuperscript{16}. As a preliminary, there is a need to distinguish between two kinds of references used in this field to cite relevant resources: 1) bibliographic references to monographs, commentaries, and journal papers (i.e. secondary sources); and 2) canonical references, namely short references that usually follow a logical rather than a physical citation schema and are used in order to cite passages of ancient texts (i.e. primary sources). Given this distinction, canonical references are of primary importance for classicists since they refer to the research object itself rather than to existing secondary sources about a certain research topic.

Hence this paper proposes a set of services that leverage several aspects of citations as they appear within journal papers: their use as navigation tools both through and between resources, their use as access points to information and thus to indexing resources, and their persistence. The following subsections will discuss the main characteristics and implications for each proposed service.
4.2 Reference Linking

Reference linking is essentially the capability in a digital environment to move directly from a citation to the resource to which it refers. Reference linking for secondary sources (i.e. modern bibliographic resources) is a feature already provided by several library catalogues, specialised search engines and e-journals. Once an interesting bibliographic reference has been found, a user can sometimes directly access it if it is referred to from an electronic resource, or they can check for the availability of an electronic copy at their own institution’s library in the case of a print resource. From a more technical perspective, the DOI technology, the CrossRef initiative and the OpenURL standard protocol have all contributed in the last few years to the solid technical foundation upon which it is now possible to build a range of reference linking services for secondary sources\textsuperscript{17}.

Reference linking to Classical primary sources (i.e. ancient texts), however, remains a desirable but not yet available service. Provided that Digital Humanities researches as a whole are an attempt to address specific scholarly users' needs with new digital tools\textsuperscript{18}, we believe that being able to browse in a digital environment the links existing between texts remains for classicists a crucial feature to be provided. When looking at the current state of digital libraries, it is clear how there is a paucity of links between ancient and modern resources. For the sake of clarity, we can consider a practical example for a moment: modern editions of ancient texts available through the Perseus Digital Library\textsuperscript{18}, Google Books and the Internet Archive (IA) are not yet linked in any way to pertinent research papers stored in the Internet Archive itself or within JSTOR.\textsuperscript{19}

For a classicist it would nevertheless be quite useful when reading a passage by Homer or Aeschylus to have at hand a list of references to resources that discussed topics concerning that passage. Or in turn, while reading a research paper or a commentary a scholar might want to read the full text of a cited passage from an ancient work.

Recently, some work has been done to address this lack of standards and tools in order to provide such a reference linking ability for primary sources. Both the solutions that were proposed up to now leverage a network protocol for the implementation, respectively OpenURL and the Canonical Text Services (CTS) protocol. The first solution is an OpenURL-based system to link canonical works in the Humanities which was identified to enhance the search interface of the \textit{Année Philologique} electronic version\textsuperscript{20}. Instead, the second solution leverages a more discipline-specific protocol to achieve the same functionality.

Indeed, the Canonical Text Services (CTS) protocol\textsuperscript{21, 22} can rightly be considered the counterpart of OpenURL for Classical primary sources. This protocol basically provides a digital equivalent to print canonical references by making text repositories accessible through logical citation schemes. Texts within a CTS repository are accessible, and thus citable, by logical hierarchical levels such as book, section paragraph or poetic line. Other recent work has also examined the technical requirements necessary to implement this feature by leveraging the CTS protocol and techniques to embed citation meta-data within (X)HTML documents\textsuperscript{23}, as well as presenting an e-journal prototype where this feature is implemented\textsuperscript{24}.

4.3 Semantic Indexing

The second feature we propose can be considered as a practical consequence of the first one. Once references are encoded for the purpose of reference linking, it becomes straightforward to leverage the available pieces of semantic information found within
citations in order to index the resources themselves. In order to allow operations such as crawling and harvesting of web resources it is necessary to envisage a solution that embeds meaningful meta-data about citations within web resources. Indexing the references to Classical primary sources found within documents provides users with an additional and significant access point to information. Indeed, while traditional search engines usually perform string-matching searches through documents, a semantic indexing system understands the semantics of citations no matter what strings are used to express them. In other words, semantic indexing allows users to access Classics e-journal contents as if they were using citations to ancient authors and works as grouping keys. In other disciplines, such as chemistry or genome research, references to chemical elements and compounds are normally extracted, classified and then employed to enhance user access to contents through the addition of semantic awareness.

4.4 Citation Persistence

In order to satisfy the observed need for persistent electronic resources and to reduce the broken link phenomenon, it is essential that Classics e-journals adopt some technical measures to deal with this problem. In particular, e-journal papers need to be citable and thus permanent and stable electronic resources, as do the external resources referred to by the paper itself.

Currently, the most suitable solution is the WebCite® system. Basically, the system acts upon user requests by caching the cited resource and providing the user with a handy link to cite it. WebCite® is part of the International Internet Preservation Consortium and its mission specifically includes long-term persistence for links and resources beyond the duration of WebCite® itself. WebCite® has also been adopted by several publishers in the Medicine field including the authoritative BioMed Central.

Another recent proposed solution, even though it differs substantially from WebCite® in terms of its declared mission, is iCyte, a commercial user application that is currently available for free. An interesting feature provided by iCyte that works with both HTML and PDF documents is the ability to refer to a specific text portion of the cited resource. When the link that is provided by iCyte is later retrieved by other users, the cited resource portion is displayed as highlighted text. Given the above noted importance of being able to create granular and precise links to electronic resources, this feature would be crucial for a persistent citation service for Classics e-journals.

5 Implementing an Innovative Classics E-journal

In this section we examine the problem of how to devise a scalable implementation of such a Classics e-journal, with a particular focus on the new proposed services (i.e. reference linking, semantic indexing). We took the whole publishing process into consideration, from the editorial workflow management to the presentation and access of research contents on the Web. Indeed, we considered the ideal scenario of a traditional print journal on the path of transition to an electronic publishing model.

5.1 Editorial Workflow Management: Choosing a Platform

As of this writing, several open source solutions exist in terms of e-journal publishing platform. Open Journal Systems (OJS), however, was by far the best fit in terms of the previously listed needs and technical requirements for a Classics e-journal. Firstly, OJS provides a plug-in mechanism that allows developers to extend its functionalities. In our case, this architectural detail allowed us to implement new value-added services tailored to
specific user needs as components that could be plugged into OJS’ overall platform.

Secondly, OJS supports batch operations (configurable through XML files) to import old journal issues. As both the case studies and our survey demonstrated, the ability to easily convert print journal back-runs into a digital format constituted a preliminary step of essential importance for a journal migrating to an electronic platform.

Moreover, the survey also illustrated that the most frequently used format for journal submissions in the field of Classics is Microsoft Word, a proprietary data format. While allowing for the fact that our data refers only to Italian journals, we can still consider this data as a reasonable indicator of tendencies in the whole field. This situation compels us to find a trade-off solution between the optimum (i.e. XML structured encoding) and the current practice (i.e. the use of Microsoft Word). Indeed, in order to provide users with advanced services we clearly need to have some kind of structured information available.

The OJS framework provides Lemon8-XML (L8X)\textsuperscript{30} for this purpose, it is a tool that converts word processor file formats into an XML mark-up compliant with the Journal Publishing Tag Set defined by the US National Library of Medicine. To reach this goal, L8X applies a set of heuristics to detect and mark up logical sections of journal papers, such as title, abstract, sections, and references among others.

5.2 Adding Meta-data (Semi-)Automatically

Since meta-data production is both a time consuming and expensive task, providing value added services in a scalable way means finding automatic procedures that can help with its creation. The meta-data we are most interested in extracting from papers is that which is needed to provide reference linking to both primary and secondary sources, namely bibliographic references and canonical references, and semantic indexing.

L8X, which is part of the OJS framework, supports integration with multiple citation parsers (Paracite, Freecite, ParsCit) that are aimed at detecting bibliographic references within scholarly papers\textsuperscript{31}. The output obtained from running those parsers is then combined with search results from freely-available online indexes in order to support manual correction of automatically detected meta-data through a graphical user interface\textsuperscript{32}. Provided that the above tools are mostly used to parse references within articles from the STM field, it is noteworthy that ParsCit was recently provided with some initial Humanities training datasets. Indeed, ParsCit employs a machine-learning approach whose main feature is the possibility of training the parser on data from different disciplines.

As far as concerns the extraction of canonical references from scholarly articles, this feature is not currently supported by any journal publishing platform. Its technical feasibility, however, was recently demonstrated, by publicly available research that is using a ParsCit-like machine learning approach\textsuperscript{33} and an open source tool called CREFEX (Canonical REFerence Extractor)\textsuperscript{34}. What we want to most strongly advocate at this point, nonetheless, is the development of a plug-in for OJS that will support the automatic extraction and parsing of discipline-specific kinds of references, such as canonical references.

5.3 Enriching Web Resources with Meta-data Embedding

Embedding meta-data within (X)HTML pages has at least two major practical results, first, it makes meta-data discoverable by both web crawlers and search engines, and, second, it can be reused by scripts or applications running on the client side. The meta-data we are particularly interested in embedding is that which concerns bibliographic
references and canonical references within scholarly journal articles. Another important reason to adopt meta-data embedding solutions is that the main search engines such as YAHOO and Google have recently started crawling semantic information embedded within web pages in the form of either Microformats and RDFa. Microformats\textsuperscript{35} and RDFa\textsuperscript{36} are currently the most suitable and widely adopted techniques used to embed meta-data within Web resources.

Some recent work conducted by the Microformats community boosted general interest about semantic meta-data embedding techniques, many of which are now regarded as a practical way to realize the Semantic Web. Low interest was expressed by the Microformats community, however, regarding the development of non-commercial Microformats. The main difference between the two technologies is that while RDFa applies to any set of RDF-encoded data, Microformats are defined and need to be approved by a developers community. Therefore, RDFa seems to be a standard more compliant with the decentralized nature of the Web itself\textsuperscript{37}.

CoinS\textsuperscript{38} is a widely adopted convention used for publishing OpenURL references in HTML and is very close to Microformats even though it was not formally approved by the Microformats community. Zotero is one of the most important applications that leverages CoinS-encoded references and it allows users to import such encoded references into their libraries. As far as concerns meta-data about canonical references, although the use of Microformats was initially suggested\textsuperscript{23,24} we now believe that the best solution would be to employ RDFa in order to embed semantic information expressed by using an ontology specifically designed for Classics-related concepts\textsuperscript{28}.

The described meta-data embedding mechanism can be added to OJS as an export filtering option to produce a (X)HTML+RDFa output, or by extending its HTML rendering mechanism.

6 Conclusions

In this paper we attempted to draw publishers' attention to a range of services that we observed to be desirable for users of the Classics field, and more generally to the importance of devising services that actually fit scholars' needs. Our hope for the immediate future is that Classics e-journals will move out of the incunabular stage by fostering value-added services specifically tailored to the needs of their users.

7 Acknowledgements

Our grateful thanks goes particularly to Prof R. Orsini and Dr F. Boschetti who supervised the research project for a master thesis in Digital Humanities this paper sprang from.

References


12. Armbruster C. Society Publishing, the Internet and Open Access: Shifting Mission-Orientiation from Content Holding to Certification and Navigation Services? [Internet].


22. Smith N. Citation in Classical Studies [Internet]. Digital Humanities Quarterly. 2009 Changing the Center of Gravity: Transforming Classical Studies Through

24. Romanello M. A semantic linking framework to provide critical value-added services for E-journals on classics. 2008;


27. iCyte [Internet]. Available from: http://www.icyte.com/


34. CREFEX (Canonical REFerences EXtractor) [Internet]. [cited 2009 Aug 31] Available from: http://code.google.com/p/crefex/


