Journal of the Text Encoding Initiative

Issue 2 | 2012
Selected Papers from the 2010 TEI Conference

CroALa

Enhancing a TEI-encoded Text Collection

Neven Jovanović

revues.org

Publisher
TEI Consortium

Electronic version
URL: http://jtei.revues.org/425
DOI: 10.4000/jtei.425
ISSN: 2162-5603

Electronic reference
Neven Jovanović, « CroALa », *Journal of the Text Encoding Initiative* [Online], Issue 2 | February 2012, Online since 03 February 2012, connection on 02 October 2016. URL : http://jtei.revues.org/425 ; DOI : 10.4000/jtei.425

CroALa

Enhancing a TEI-encoded Text Collection

Neven Jovanović

AUTHOR'S NOTE

The author would like to thank Mark Olsen (ARTFL Project, University of Chicago) for his helpful correspondence on PhiloLogic.

1. Introduction

1 *Croatiae auctores Latini* (CroALa) is a digital collection intended to provide open access to and computational manipulation of as many texts of Croatian neo-Latin literature as possible (Jovanović et al. 2009). This literature, created by Croatian authors writing in a 'cultural', learned language, is a phenomenon of surprisingly long endurance, lasting at least nine centuries: from around 1100 to the present day.

2 CroALa, published at the University of Zagreb under a Creative Commons license, collects peer-reviewed digital editions of texts encoded in TEI XML, which can be retrieved and searched over the Internet through an implementation of the PhiloLogic system.[1] The collection also includes secondary prosopographical and bibliographical data on Croatian authors and texts in Latin. In February 2011 CroALa comprised 106 documents by 75 authors.

3 Access to primary texts of Croatian neo-Latin—which were until now widely dispersed in old or local editions, varying in editorial quality and approaches—is something that the scholarly community would immediately understand and appreciate. However, before CroALA the community had yet to explore the heuristic and interpretational value of computational manipulation, such as the complex bibliographic and structural querying which PhiloLogic makes possible.[2] Therefore, CroALa provides support for such querying,

using informative, corrected, and strategically placed metadata as well as examples and explanation of searches.

4 Potential users of CroALa include researchers for whom reading Latin is a secondary skill (such as historians) and students who are just beginning to master the language. For the collection to be used as widely as possible (the number of users will in any case remain modest), we needed to provide good language support.

5 Finally, searching and retrieving are not the only reasons to prepare a TEI XML encoded collection of texts. Croatian Latin will be better preserved digitally and will be more likely studied if its texts are published in different places and in different formats, exploiting the full potential of an open licence and a transformable markup language.

6 Here we present six ways in which CroALa enhances support for complex querying for users with less Latin language competence and different interests, and for wider dissemination of Croatian Latin texts.

2. Fine-tuning Metadata and PhiloLogic

7 The practice of text encoding usually distinguishes between data on the one hand and programs and interface on the other. As Deegan (2006) writes, 'however important the programs used to create and deliver the edition and the interface through which it is accessed, scholars must always remember that these are likely to be the least durable part of any electronic edition.' In experimenting with CroALa, we have taken a slightly different path. We started with very basic bibliographic and structural encoding and then explored how interesting queries of the data could be formed in PhiloLogic. It helped that both the tool and the encoding scheme were designed to be modifiable. The authors of PhiloLogic recommend using TEI Lite, but the system 'is known to handle more than [that]' (Philologic 2010), and we were indeed able to include elements from other TEI sets without difficulty.

8 CroALa was started with the general aim of making Croatian Latin texts accessible and searchable. We did not have a clear idea exactly *what* we wanted to search for, so we did not restrict the collection to just one set of research problems. That is, we did not select only literary texts, or texts with a specific theme or genre, or texts from a specific period. After some time, having encoded enough texts and tested different approaches, common trends began to emerge: quite elementary bibliographic, textual, and lexical data turned out to relate to each other in interesting ways. CroALa's interface, through PhiloLogic's forms, enables querying both strings and metadata fields—titles, authors' names and dates, dates of creation and publication, genres, historical periods, etc.

9 To ensure maximum compatibility of internal (data) and external (program and interface) structure, two sets of constraints had to be made, one on the document encoding schema and another on scripts which control the search procedure. The TEI schema—especially the TEI header—was customized, first using the Roma interface, then finishing by hand—to exclude all unnecessary elements and make sure that PhiloLogic loaders (Perl scripts which load bibliographic data into indices and MySQL tables) will find everything they need while the documents remain valid TEI XML.[3] The loaders were adapted and constrained, e.g., to expect in the `<sourceDesc>` element only `<biblStruct>`, and nothing else from the model.biblLike model class. Such two-way constraints keep CroALa documents valid TEI XML and ensure that PhiloLogic, as our current full text search

system, is used to its full potential. As all changes and adjustments are documented, encoders can follow the rules better, and administrator tasks become less demanding.

3. Providing Language Tools for Users

10 We mentioned that CroALa should serve the needs not only of Latin and neo-Latin scholars but also of other readers and users, such as students or scholars from other areas of research who might well need help with Latin words. They can get it in two ways: through the PhiloLogic interface (on the server side) and through the Alpheios Reading Tools (on the client/browser side).

11 PhiloLogic has a function called QuickDict, which allows the user to select a word and perform a dictionary lookup. Since Latin is an inflected language, looking up a word is a two-step process: the selection is sent first to a parser, which returns the lemma, and then a dictionary entry is retrieved. For both tasks our PhiloLogic installation uses an external web service: the "Perseus under Philologic" site at the University of Chicago.[4]

12 A new and interesting client-side application for reading texts in Greek and Latin is Alpheios Reading Tools, which are Firefox extensions for reading and learning languages released in beta in 2010. When Alpheios is installed and a reader double-clicks a word, Alpheios parses it and provides a short dictionary definition. Developers have also provided a version as a web service (Alpheios Project, Ltd 2011), which can be enabled in a webpage by simply adding a single element to the HTML document header. We added this to the default TEI-to-HTML XSLT stylesheet and to the CroALa PhiloLogic results page template.

13 PhiloLogic's QuickDict and the Alpheios Reading Tools contain different features. QuickDict is not browser-dependent, but Latin parsers currently available online (which QuickDict uses) do not take into consideration orthographical variants that have arisen over two thousand years of writing in Latin. Alpheios fares better here, but it is bound to Firefox. Both sets of tools offer only English vocabulary translations, which may considerably lower its value for classroom and study use in Croatia and the wider Central European region.

4. Adding Supplemental Material and Links

14 A subset of the intended users of CroALa belongs to a scholarly community that is accustomed to reading old books. With this in mind the editors decided to follow the practice of an important German neo-Latin digital library, the DFG-Projekt CAMENA, Heidelberg-Mannheim. CAMENA publishes both machine-readable texts and digital images of old books. This is an incremental strategy, providing quick access to works whose full digitisation involves a costly and time-consuming transcription and encoding —a responsible strategy for treatment of cultural heritage. In the words of the editors at CAMENA: 'die den Bildseiten beigegebenen maschinenlesbaren Texte sollen nicht die zeitgenössischen Ausgaben in den Hintergrund drängen' ['the machine-readable texts which accompany the page images should not replace the historical editions'] (CAMENA 2011).

15 By including digital images in the collection, CroALa meets another need of Croatian neo-Latin scholarship: the need to 'collect and connect' valuable material freely accessible yet

widely dispersed on the Internet. Croatian neo-Latin texts are made accessible by other digitalization projects, especially those undertaken by libraries such as the Bibliothèque nationale de France or Bayerische Staatsbibliothek or by Google.

16 The basic tool in collecting such material is, of course, a bibliography or a catalogue of Internet resources. Experience shows that each resource and collection approaches digitization in a different way, so that even basic navigation—e. g., non-linear skipping from the table of contents to beginning of a chapter—can turn into a non-trivial task, not to mention the challenge of finding an appropriate hyperlink to cite as a pointer to such a place in a digitized book.

17 In order to make documents digitized and published elsewhere navigable in a uniform way, CroALa prepares bare-bones TEI XML documents consisting primarily of links to digital page images (given as `@facs` attributes in `<pb>` elements), supplemented by bibliographic data (in the TEI header), titles (if available), and incipits and explicits of each page.[5]

18 Another useful addition to CroALa documents are links to digital versions of texts quoted by authors whose texts are found in CroALa. This is especially important in the case of commentaries, such as the one on Catullus by Marko Marulić, but also in religious and theological texts, which rely heavily on biblical and patristic literature.

19 Linking to external resources poses two problems. First, where to link to in the case of citations to canonical works without reference to a particular edition, or in the case of a work digitised several times in several different locations, and what to do if the link gets broken if the resource moves, changes, or goes away? Second, where and how to encode such a link?

20 We were quite aware that there is no one authoritative and universal resource on the Internet to link to for canonical works like those of classical authors or the Bible; moreover, we needed a resource which will be freely accessible and as reliable and fast as possible. Also, the resource should support linking: it should provide a uniform and logical way to construct or harvest a link to a particular passage.

21 For the time being, quotes in CroALa from classical Greek and Latin authors, as well as the Vulgate, link to "Perseus under PhiloLogic". These links were encoded as `<ref>` elements with `@target` attributes, which PhiloLogic renders as hyperlinks.[6]

5. Bibliographic Links and Citations

22 Citing is a basic scholarly activity; therefore, a resource, if it deserves to be cited, will enhance its prominence—and, indirectly, its credibility—by being easily citable. But although online and digital texts have been an essential part of humanities research in recent years, footnotes and bibliographies found in recent neo-Latin scholarly work show that our colleagues do not feel completely at ease in citing digital sources. Printed books and articles (or their digital facsimiles, silently taken for 'the real thing') may seem somehow simpler to describe bibliographically. That is why we wanted to help users cite CroALa.

23 We tried BibSonomy, a web-based reference management and publication-sharing system. A chicklet for BibSonomy was embedded in footers of pages served by PhiloLogic,[7] enabling users with a BibSonomy account to create bookmarks of the collection, of their

search results, or of individual texts. Those without an account can quickly and freely create one.

24 Citing CroALa would be even easier if the user could create not only a BibSonomy bookmark but also a publication record. (BibSonomy has a separate category for this, built on publication formats of BibTeX.) To achieve this, it will be necessary to write a BibSonomy scraper for CroALa and contribute it to the service.[8]

25 CroALa texts are not only themselves citable but also contain citations to other works. The bibliographic records entered in `<sourceDesc>` of our documents could be presented by PhiloLogic as 'live' hyperlinks when they refer to a book which has a digital facsimile in a repository such as the Internet Archive or the Bayerische Staatsbibliothek, regardless of whether individual page facsimiles are referred to in the body of the CroALa document (see above). PhiloLogic provides two points where such hyperlinks could be implemented: in the display of the TEI header (currently presented inside an HTML `<pre>` element) or in the bibliographic display of PhiloLogic (formatted in the `mkBiblio` subroutine of `philosubs.pl`).

6. Digitizing and Providing Access to Bibliographies and Prosopographies

26 Scholarly research of Croatian neo-Latin relies not only on primary texts but also on bibliographies, prosopographies, and similar finding aids. Querying such resources digitally can itself lead to valuable insights. That is why CroALa contains digitized versions of a prosopography of important Croatian Latin writers and a bibliography of their works (at present, a list of printed titles published before 1850), both encoded in TEI. These are kept separate from the primary CroALa documents, but here too we use PhiloLogic for online display and search.

27 Our finding aids are essentially lists; the prosopography is inside a `<listPerson>` element, and the bibliography inside a `<listBibl>`. Because PhiloLogic by default recognizes only `<div>` and a selection of TEI subdiv elements, we created an ad hoc XSL transformation for bibliographical and prosopographical lists to allow them to be used in PhiloLogic,[9] keeping, of course, the original TEI documents as master copies.

7. Disseminating the Texts in Outside Repositories

28 One of the aims of digitizing texts by Croatian Latin writers is to attract more readers and potential researchers. To achieve this, we decided, first, to publish all texts under an open license (Creative Commons BY-NC-SA), and second, to disseminate our primary texts as widely as possible, sending them to popular open repositories.

29 On the Internet there seem to be two main ports of call for Latinists who want to contribute texts; the Latin Library and Vicifons, the Latin-language version of Wikisource. The Latin Library, which every student of Latin encounters sooner or later, accepts HTML documents with minimal formatting, so CroALa texts can be published there quite easily if transformed by the default set of stylesheets prepared by TEI and released on Sourceforge. The Latin Library remains intentionally discreet about its sources,[10] has relatively little visible concern for standards and preservation, offers just elementary

navigation, and gives the user practically no means of fulltext search. These tasks are better carried out by Vicifons, which relies on Wikipedia's infrastructure and follows the general publishing philosophy of Wikimedia Foundation. The main downside to Vicifons is that it is younger than the Latin Library[11] and therefore less known.

30 Vicifons uses Wiki markup syntax, so we have prepared several XSLT stylesheets to transform the TEI CroALa documents into the Wiki format, using three Vicifons templates (predefined instructions which duplicate the same content across more than one wiki page) for identifying a work (`<Titulus2>`) for navigation (`<Liber>`), and for metadata (`<OperisInfo>`). As test cases, we published a longer poem, and a collection of short texts. First versions of Vicifons XSLT stylesheets—one for prose, one for poetry, one for bibliographic data—were published on TEI Wiki.

31 Another accessible and visible open repository is the Internet Archive, where texts in the public domain can be uploaded (PDF format is recommended) with a bibliographic description for distribution and archiving. After some experiments, we decided that the texts to be published in the Internet Archive will be transformed first to HTML (using TEI stylesheets) and then to PDF.

8. Conclusion and Future Plans

32 Having completed the initial phase of developing the *Croatiae auctores Latini* digital collection, we are currently trying to enhance its research potential. Two main aims guide us. Firstly, we want to make Croatian texts in neo-Latin easily accessible and searchable in novel ways. Secondly, we want to attract the interest of scholars from Croatia and abroad, from neo-Latin studies and other disciplines. Accordingly, all configurations and modifications proposed here are modelled on standard approaches to neo-Latin literature. We neo-Latin scholars think in terms of genres, periods, authors, and works; therefore, CroALa needs to provide such metadata. We research an author, a period, a theme, and a region, and therefore the search and retrieval system must be able to use these categories for creating (and saving) 'ad hoc corpora [and] dynamic sub-collections of texts that can be automatically compiled in response to a query' (Bamman and Smith 2010). While we go to libraries looking for physical copies of old, difficult-to-get, and difficult-to-read editions, if somebody offers us an online digital facsimile of such an edition, this is also useful. We also rely on bibliographies and lexicons, which means that lists of titles and words are welcome offerings in CroALa.

33 It turned out that practically all we wanted can be achieved by combining XSL-transformable TEI XML with the extensively configurable rules PhiloLogic has for conversion of original encoding to HTML, and for accessing external language tools. We came to consider this combination as a two-way approach, a half-way meeting. On the one hand, the TEI scheme is constrained to take advantage of PhiloLogic; on the other hand, PhiloLogic is configured to better search and retrieve such constrained documents.

34 The CroALa collection relies significantly on other Internet resources and services. Our links point towards various repositories of digital facsimiles: words from our texts are parsed and lemmatized externally, and users are offered an option to store citations from CroALa on another web-based reference and publication-sharing system. This reliance on other services was a conscious decision, made both because of our limited resources and because we believe that interoperability is important for this stage of digital humanities.

35 What needs to be explored further is how PhiloLogic can make use of the TEI's module for analysis and interpretation. Up to now, our thematic collections (e.g., of texts praising Dalmatian cities, *Laudationes urbium Dalmaticarum*) remain isolated from the main CroALa corpus, in which texts are described not by themes, but by authors, periods, and genres. A mechanism for adding into this mix themes and motives, however woolly such categories, would enable CroALa and its users to better model interpretation of Croatian neo-Latin, exploring different implications of such modelling.

BIBLIOGRAPHY

Alpheios Project, Ltd. 2011. "HTML Integration - Auto Enabling Alpheios." Accessed July 24. http://alpheios.net/content/html-integration-auto-enabling-alpheios.

Bamman, David and David Smith. 2011. "Extracting Two Thousand Years of Latin from a Million Book Library". Preprint. http://www.perseus.tufts.edu/~amahoney/01-jocch-bamman.pdf.

BibSonomy. 2011. "All sites supported by our scrapers." Accessed October 22. http://www.bibsonomy.org/scraperinfo.

CAMENA. 2011. "CAMENA - Lateinische Texte der Frühen Neuzeit." Accessed October 22. http://www.uni-mannheim.de/mateo/camenahtdocs/camena.html.

Deegan, Marilyn. 2006. "Electronic Textual Editing: Collection and Preservation of an Electronic Edition." *Electronic Textual Editing*. Edited by John Unsworth, Katherine O'Brien O'Keeffe, and Lou Burnard. New York: Modern Language Association of America. Preprint. http://www.tei-c.org/About/Archive_new/ETE/Preview/mcgovern.xml.

Jovanović, Neven, et al., eds. 2009. *Croatiae auctores Latini.* Zagreb: Faculty of Humanities and Social Sciences, University of Zagreb. http://www.ffzg.unizg.hr/klafil/croala.

Latin Library. 2011. "About These Texts". Accessed February 15. http://www.thelatinlibrary.com/about.html.

Perseus under PhiloLogic. 2010. "About Perseus under PhiloLogic.". http://perseus.uchicago.edu/about.html.

Philologic. 2010. "What is PhiloLogic?" Accessed August 24. http://sites.google.com/site/philologic3/.

Sutton, Dana F. 2011. *Analytic Bibliography of On-line Neo-Latin Texts.* Irvine: The University of California. http://www.philological.bham.ac.uk/bibliography/

Wikipedia contributors. 2011. "Wikisource." *Wikipedia, The Free Encyclopedia.* Last modified February 8. http://en.wikipedia.org/w/index.php?title=Wikisource&oldid=412690599.

NOTES

1. PhiloLogic is a full-text search, retrieval and analysis tool developed by the Project for American and French Research on the Treasury of the French Language (ARTFL) and the Digital Library Development Center (DLDC) at the University of Chicago. The tool exists in a Free Software, open source implementation (which was used for CroALa), and supports large TEI-Lite document collections 'out of the box'; other varieties of encoding can be searched and served with some adaptation to the system, which was designed to be modifiable. The structure is modular: 'a textbase is treated as a set of coordinated or related databases, typically including an object (units of text such as a letter, scene, document, etc) database, a word forms database, a word concordance index mapped to textual objects, and an object manager mapping text objects to byte offsets in data files. Each of these databases is stored and managed using its own subsystem.' (Philologic 2010).

2. A good illustration for such communal attitude would be the most important digital neo-Latin reference tool (Sutton 2011). Created in 1999 and listing about 38,470 records in June 2011, this monumental internet bibliography offers only basic search capabilities: an index by authors and a Google custom search.

3. We followed the TEI specifications offered by PhiloLogic.

4. An important collection of classical Greek and Latin texts is encoded in TEI XML and published under an open license by the Perseus Project at Tufts University. As the name itself suggests, "Perseus under PhiloLogic" is a different deployment of the same texts, which are searched and retrieved by the system for CroALa.

5. An example of such a document and its encoding can currently be seen in the Slike test database. Although we generally follow PhiloLogic specifications, we felt we needed to define two different kinds of facsimile links: pointers to local images (which system constructs links by default) are encoded as `@facs` in a `<milestone>` element; and external links, which are given as `@facs` in `<pb>`.

6. The necessary modifications to `philosubs.pl` can be found at our PhiloLogic configurations pages.

7. See the Javascript code at the Bookmarklet buttons for BibSonomy page, and the application of it at any CroALa page.

8. On BibSonomy and scrapers, see BibSonomy 2011, and the entry "Metadata Scraping Service", BibSonomy blog: news about www.bibsonomy.org, November 11, 2008.

9. We wrap each `person` and `biblStruct` element in a wrapper with some additions for easier navigation within PhiloLogic search results; see the XSLT stylesheet at the ProsopToDiv.xsl page of the TEI Wiki.

10. The almost anonymous editor of the Latin Library collects Latin texts in the public domain, but does not indicate the exact origin of a particular contribution, stating simply that 'the texts are not intended for research purposes nor as substitutes for critical editions' (Latin Library 2011).

11. While the Latin Library goes back to 1998, Wikisource in Latin was created in August 2005 (Wikipedia contributors 2011, note 2).

ABSTRACTS

Croatiae auctores Latini (CroALa), a text collection first published in 2009, makes freely accessible Latin texts written by or about people of Croatian origin from the Middle Ages to the twentieth century. The collection is intended primarily for scholars of Latin and neo-Latin literature and language and of Croatian history and culture. Texts are encoded in TEI XML and made searchable and readable online using PhiloLogic.

This edition is intended to become a starting point and a useful tool for serious research, both traditional and digital. What can be done, besides adding more texts, to stimulate such research? This article describes the editors' solutions in the following six areas:

- fine-tuning the metadata and user interface to enable grouping of texts by author, period, genre, theme;
- providing language tools for users;
- adding supplemental material and links;
- making citations of CroALa texts easier and providing hyperlinks between texts;
- digitizing and providing access to finding aids for effective search and retrieval;
- disseminating the texts in outside repositories.

These enhancements are realized by XML encoding, XSLT stylesheets, and configuring PhiloLogic. These solutions are simple to implement and are within reach of an advanced, non-programmer computer user. The documentation of these solutions serves a twofold pedagogical purpose: first, to help newcomers find their way around the immense potential enabled by the TEI encoding scheme and, second, to encourage other humanities scholars to become makers—or, at least, adapters—of TEI-based digital tools. Simply put, if I did it, you can do it too!

INDEX

Keywords: citation, Croatian neo-Latin literature, interoperability, language tools, metadata, PhiloLogic, text dissemination

AUTHOR

NEVEN JOVANOVIĆ

Neven Jovanović was born 1968 in Zagreb, Croatia, and received his PhD in classical philology at the University of Zagreb in 2005. From October 2006 he has been a University Docent (Adjunct Professor) at the Department of Classical Philology, Faculty of Humanities and Social Sciences, University of Zagreb. From 2008 Jovanović has been one of the editors of Colloquia Maruliana, an annual on Croatian humanist and renaissance literature (published in Split); from 2009 he has been the editor of Croatiae auctores Latini, a collection of digital texts by Croatian Latin writers. In 2010 he won a Google Digital Humanities Research Award for a project proposal A Profile of Croatian neo-Latin. His main research interest is Croatian neo-Latin literature, both as a specific aspect of classical tradition and reception and as a digital humanities research theme.