The workspace for collaborative editing

Houghton, H.A.G.; Smith, Catherine; Sievers, Martin

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Houghton, HAG, Smith, C & Sievers, M 2014, The workspace for collaborative editing. in Digital Humanities 2014 Conference Abstracts, EPFL–UNIL, Lausanne, Switzerland, 8–12 July 2014. Lausanne, pp. 210-1, Digital Humanities Annual Meeting, Lausanne 2014, Lausanne, Switzerland, 8/07/14. <http://dharchive.org/paper/DH2014/Paper-224.xml>

Link to publication on Research at Birmingham portal

digital humanities
Lausanne – Switzerland '14
UNIL | Université de Lausanne
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

THE WORKSPACE FOR COLLABORATIVE EDITING

Houghton, Hugh
Institute for Textual Scholarship and Electronic Editing (ITSEE), University of Birmingham

Sievers, Martin
Trier Center for Digital Humanities (Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften, Universität Trier)

Smith, Catherine
Institute for Textual Scholarship and Electronic Editing (ITSEE), University of Birmingham

Category:	Long Paper
Session:	1
Date:	2014-07-09
Time:	09:00:00
Room:	414 - Amphimax

The Workspace for Collaborative Editing is a project funded by the AHRC (UK) and DFG (Germany) between September 2010 and December 2013. It has the goal of creating an online workspace to support the production of the *Editio Critica Maior*of the Greek New Testament by teams based in Birmingham, Münster and Wuppertal and collaborators dispersed all over the world.[1] The edition has been in progress since the 1990s, but the obsolescence of key tools and encodings have led to this ambitious project to connect all the different stages of the editorial process through online interfaces and shared databases.

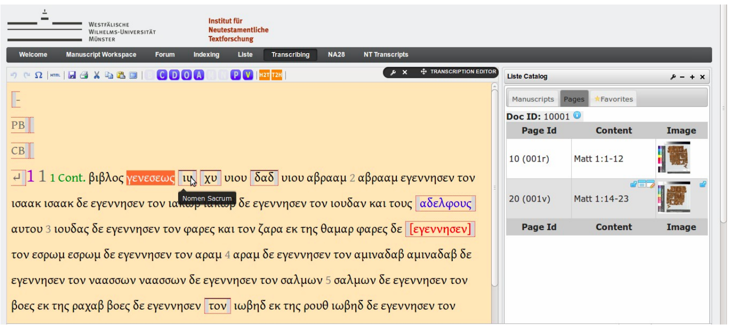
The production of a critical edition involves the identification and selection of manuscripts to be included, the acquisition of images, the creation of full-text electronic transcriptions (which are themselves published as separate electronic editions, linked to the electronic apparatus, enabling further research in related fields), the automatic comparison of these transcriptions to generate a critical apparatus of all variant readings, the editing of this apparatus by scholarly editors to filter out 'noise' and prepare the data for analysis using genealogical tools, the addition of evidence from early translations and biblical quotations and the publication of the material in electronic and printed form.

The aim of the Workspace project has been to adopt existing standards and open-source solutions in order to create a lightweight architecture capable of being easily renewed and updated, so that both the data and software created may be reused by other projects. The result is an open-source browser-based environment written in Python and Javascript. The core software consists of a MongoDB database and the asynchronous web application framework MAGPY. Data is stored in JSON and made available via a RESTful interface. On top of this is a layer of applications which call the relevant data objects for the individual editing processes. The goal of transparency at every level of the editing process means that a record is kept of each object at each stage of the process, and any modifications introduced are treated as additional records rather than replacing existing data.[2]

The Greek New Testament provides a very specific use-case, with a large amount of data already created and highly developed editorial principles. In addition, ongoing work by existing editorial teams offers the opportunity for immediate testing in real-life situations. Developing in these circumstances can be a challenge, with the evolution of guidelines, changes of editorial practice and 'creeping featurism'. The system needed to make existing legacy data compatible with the much more detailed XML encoding developed by the project and cater for as many known and potential scenarios as practicable. The dispersed team of editors was often called upon to codify their procedures and reach a common mind on problems presented by live data, including agreeing changes in policy. As a result, the creation of the Workspace has proceeded hand in hand with the development of different stages of the edition as a whole.

The two principal areas in which the Workspace meets a pressing need are the development of a transcription editor, which produces and allows the editing of valid XML in a WYSIWYG environment, and a collation editor which enables the scholarly creation of a critical apparatus. Both of these are browser-based, in order to enable dispersed collaborators to work with differing operating systems and contribute directly to the central data store.

The Transcription Editor has been created by team members at the Trier Center for Digital Humanities and released as open-source at the end of the project.[3] Its basis is the platform independent TinyMCE package.[4] A set of options for mark-up was then developed through a series of menus and shortcuts (cf. Figure 1). The aim is to allow student and volunteer transcribers not familiar with XML to work in an environment which matches as closely as possible the format of the transcriptions already published in the system. The mark-up in the browser uses HTML encoding. An export function converts this into XML matching the specifications developed by the project.[5] Likewise, an import function is required in order to support the editing of existing transcriptions. Some of the problems include the encoding and display of paratextual information, normally located in the margins of a manuscript. The dialogue box for entering this information has to have the same functionality as the main transcription interface for recording unclear or supplied text, corrections and so on. The concept has therefore been developed of the "editor-within-an-editor" which makes this possible. A problem with the import of existing transcriptions is the sequence within which elements were nested within the XML. As a result, it has been necessary to establish a system of tag sequences supported by the editor. The standalone nature of the Transcription Editor and its use of an agreed set of TEI encoding means that it can be installed as a plug-in to different environments, including the New Testament Virtual Manuscript Room (NTVMR 2.0)[6] as well as the Workspace for the production of the critical edition.



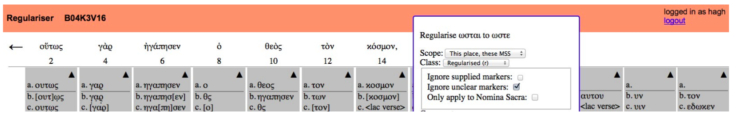
1 1 1 Cont. βιβλος γενεσεως ιυ χυ υιου δαδ υιου αβρααμ 2 αβρααμ εγεννησεν τον ισαακ ισαακ δε εγεννησεν τον ιακωβ ιακωβ δε εγεννησεν τον ιουδαν και τους αδελφους αυτου 3 ιουδας δε εγεννησεν τον φαρες και τον ζαρα εκ της θαμαρ φαρες δε [εγεννησεν] τον εσρωμ εσρωμ δε εγεννησεν τον αραμ 4 αραμ δε εγεννησεν τον αμιναδαβ αμιναδαβ δε εγεννησεν τον ναασσων ναασσων δε εγεννησεν τον σαλμων 5 σαλμων δε εγεννησεν τον βοες εκ της ραχαβ βοες δε εγεννησεν τον ιωβηδ εκ της ρουθ ιωβηδ δε εγεννησεν τον

Doc ID: 10001

Page Id	Content	Image
10 (001r)	Matt 1:1-12	
20 (001v)	Matt 1:14-23	
Page Id	Content	Image

Fig. 1: The Transcription Editor in the NTVMR environment. Based on the selection different menu options for breaks, corrections, deficiency, ornamentation, abbreviations, marginals, notes and punctuation are offered. Mouseovers and different colours help the users to identify different structures.

The Collation Editor provides an interface to the CollateX engine developed by the INTEREDITION project, the successor to the COLLATE program by Peter Robinson.[7] This software performs one of the most mechanical and error-prone tasks in an edition, namely the comparison of all witnesses in each variation unit to build up a critical apparatus. Each file is aligned using an algorithm taking into account not just spelling variations, additions, omissions and substitutions, but also transpositions within each block of text. However, the output still requires considerable input from scholars in order to clean up the raw data for publication as a critical apparatus. The first stage is regularisation, the elimination of insignificant variations such as spelling errors. The variant readings are set out underneath a base text, with the witnesses attesting each reading visible in a mouseover box (see Figure 2). An interface built using the redips drag-and-drop library allows editors to drag-and-drop the readings for regularisation onto the correct form.[8] For each regularisation, a dialogue box requires users to state the scope of the regularisation and also its nature. Once this is completed, the regularisation is marked in grey and a rule is saved to the database. The 'recollate' button sends the data back through CollateX, preferring the regularised token to the original form where present. This means that a different configuration of readings may appear in each column, as the data is cleaned up and a better match is made by the collation algorithm. The second stage involves setting the length of each variant unit, again implemented through a user-friendly drag-and-drop interface for combining or splitting neighbouring columns. One of the dangers with this interface is changing the overall sequence of words in a manuscript by combining different units and repositioning readings. A checking mechanism has therefore been developed which warns the user as soon as the sequence of any manuscript has been disrupted. On some occasions, the data is best displayed as two units of different lengths. By right-clicking on the relevant reading, it can be sent to a line below as an "overlapping variant", which can then be combined and manipulated like the other columns. One further complication is that an overlapping variant such as a lengthy transposition of words may also contain a reading which should cited in the main sequence. The system therefore makes it possible to duplicate such readings. The final stage is the ordering of variant readings within each unit and assigning the appropriate reading identifier. From here, the apparatus can be output in a number of forms, such as a positive or negative plain text apparatus, an XML encoded apparatus, or a set of values for incorporation into a database for phylogenetic analysis. The information added in the regularisation dialogue box makes it possible to generate automatically the lists of original forms for orthographic variants and erroneous readings which are printed in an Appendix in the *Editio Critica Maior.*



οὕτως	γὰρ	ἠγάπησεν	ὁ	θεὸς	τὸν	κόσμον,
2	4	6	8	10	12	14
a. ουτως b. [ουτ]ως c. ουτως	a. γαρ b. γαρ c. [γαρ]	a. ηγαπησεν b. ηγαπησ[εν] c. ηγα[πη]σεν	a. ο b. θς c. [ο]	a. θεος b. ηγαπησεν c. θς	a. τον b. των c. [τον]	a. κοσμον b. [κοσμον] c. <lac verse>

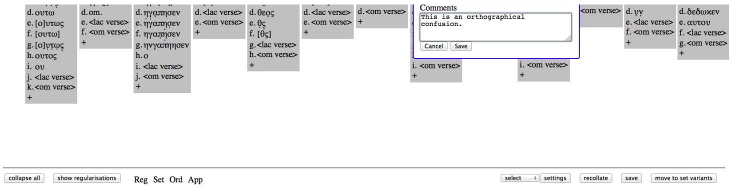
Regularise ωσται to ωστε

Scope: This place, these MSS

Class: Regularised (r)

Ignore supplied markers: □
Ignore unclear markers: ☑
Only apply to Nomina Sacra: □

	a.	a.
αυτου	b. υν	b. τον
<lac verse>	c. υιν	c. εδωκεν



This is an orthographical confusion.

Fig. 2: The regularisation interface with the dialogue box displayed.

The presentation will briefly demonstrate the Workspace, especially the two interfaces described above. We will discuss some of the problems encountered during its development, along with their solutions. Although the scope of the original project was specifically to support an edition of the Greek New Testament, a pilot project to customise the environment for an edition of Avestan texts will be outlined: from here, we hope that it will be possible to develop the Workspace for use with other textual traditions.

References

For the history of the ECM project, see **Klaus Wachtel**, *Editing the Greek New Testament on the Threshold of the Twenty-first Century" Literary and Linguistic Computing 15* (2000), 43–50; **D.C. Parker and Klaus Wachtel**, *The Joint IGNTP/INTF Editio Critica Maior of the Gospel of John: its goals and their significance for New Testament scholarship*, presented at the Annual Meeting of SNTS, August 2-6, 2005, Halle. epapers.bham.ac.uk/754/.

2. For documentation, see zeth.github.io/magpy/index.html. The source code may be downloaded from github.com/zeth/magpy.

3. sourceforge.net/projects/wfce-ote/

4. www.tinymce.com/

5. **H.A.G. Houghton**, *The Electronic Scriptorium: Markup for New Testament Manuscripts*, in Claire Clivaz, Andrew Gregory and David Hamidovic (edd.), Digital Humanities in Biblical, Early Jewish and Early Christian Studies, Leiden: Brill (2013), pp. 31–60; the latest version of the specifications is at epapers.bham.ac.uk/1727.

6. ntvmr.uni-muenster.de/en_GB/transcribing

7. **P.M.W. Robinson** (1994), *Collate: Interactive Collation of Large Textual Traditions, Version 2. Computer Program distributed by the Oxford University Centre for Humanities Computing*, Oxford. The history of Collate is described by Robinson in a 2007 blog post at: www.sd-editions.com/blog/?p=15. For CollateX, see www.interedition.eu/. The source code is available from collatex.net/.

8. www.redips.net/javascript/drag-and-drop-table-content/