2007-11-21

Plan for an Improved Infrastructure for the Encyclopedia of Diderot and d'Alembert: Collaborative Translation Project

Hawkins, Kevin

SCHOLARLY PUBLISHING OFFICE WHITEPAPER[*]

Plan for an Improved Infrastructure for the *Encyclopedia of Diderot and d'Alembert: Collaborative Translation Project*[1]

Kevin S. Hawkins

Executive Summary

The Scholarly Publishing Office (SPO) of the University of Michigan University Library proposes moving the *Encyclopedia of Diderot and d'Alembert: Collaborative Translation Project* from the DLXS digital library suite, hosted by the University Library and published by the SPO, to a wiki hosted by SPO at scholarlypublishing.org (SPO's domain name for experimental publishing projects). SPO feels that the time is right for the Project to take advantage of technology that supports a more flexible and innovative model for translation by scholars than allowed with the current system. The specific wiki software used is to be determined. A prototype will be built (Phase I), and if it meets the approval of the Project Directors, a full migration will be undertaken (Phase II).

SPO envisions a wiki for the Project that is publicly readable but editable only by Volunteers, the Project Directors, and SPO staff. The Project Directors would still screen Volunteers, adding them to a list of users authorized to edit. All edits would be tracked and recorded for each user, as with all wikis.

Each article in the Project would constitute a page in the wiki. All users adding or editing content on the site would be prompted at the point of creation or editing of a page to agree to the terms of a license for redistribution of the content.

Articles could be monitored for changes by the Project Directors using the *recent changes* page of the wiki software. Others besides the Project Directors could share the burden of verifying changes. In addition, any contributor could choose to watch pages of interest—most likely articles they originally translated but possibly also others.

The system of assigning articles for translation ahead of time could be maintained, or the model could be adjusted somewhat to fit a more common wiki model. Pages in the wiki can have one or more *categories* assigned to them, allowing easy browsing by category.

Despite the challenges, moving the Project to a wiki should allow the Project to grow more naturally and allow the Volunteers, Project Directors, and SPO staff to experiment with new models of scholarship. It should also save the Project Directors time spent on the Project, allowing the community of Volunteers to share some of the burden of stewarding the Project.

Background

The *Encyclopedia of Diderot and d'Alembert: Collaborative Translation Project* (*the Project*) was launched in October 2002, after a call for participation in January of that year resulting in a modest initial collection of 21 articles translated from the first edition of the *Encyclopédie* (published in French between 1751 and 1777) to English. The Project now

[1] This whitepaper is an expanded and revised version of a previous Scholarly Publishing Office Whitpaper, "Proposal for an Improved Infrastructure for the *Encyclopedia of Diderot and d'Alembert: Collaborative Translation Project*," available at http://www.ultraslavonic.info/preprints/20070112.pdf.

includes 500 articles (out of more than 70,000 in the original *Encyclopédie*), with a new batch added to the Project's website nearly every month. The Project is published by the Scholarly Publishing Office[2] (*SPO*) of the University of Michigan (U-M) University Library. There are three Project Directors—Dena Goodman, Jen Popiel, and Bryan Skib—and various SPO staff members have worked with the Project over time.

Current system

The project is *collaborative* in that anyone can contact the Project Directors and volunteer to contribute to the translation of the *Encyclopédie*. Volunteer translators (*Volunteers*) email the Project Directors, and Dena screens applicants based on their qualifications; unqualified Volunteers are not accepted. Volunteers either choose particular articles to translate from the original *Encyclopédie* or are assigned them based on their subject interests.

Each article is assigned for translation by Jen, who tracks this assignment and offers the Volunteer the French text copied from the ARTFL Encyclopédie Project, which presents the full French text of the *Encyclopédie* online.[3] In cases where a Volunteer fails to translate the article in a timely fashion and a second Volunteer volunteers to translate the same article, the Project Directors occasionally impose a deadline on the first Volunteer, after which the second Volunteer's translation of the article will be accepted instead.

Volunteers email their translated articles as word processor files to Jen, who edits them as needed for conformance to the Project styleguide. Each translated article emailed to Jen must be accompanied by a license (a Word document "signed" by the contributor) giving the Project the right to publish the work online and giving users of the site the right to copy translations for personal and academic purposes.[4]

Jen emails the translated articles as Word documents or RTF files to SPO on approximately the 15th of the month for publishing by the end of the month. She is responsible for only sending articles which she has received a license to publish.

Metadata for articles in progress is stored in a FileMaker database (the *Header Database*) accessible to the Project Directors and to SPO staff. Jen enters the data initially, and SPO staff revises it as necessary when publishing new articles.

The Scholarly Publishing Office uses the DLXS digital library suite (*DLXS*) to deliver the Project, as it does with nearly all its other publications. A translated article, not a page of text of either the original *Encyclopédie* or of the translated version, constitutes the unit of electronic text. The electronic text contains hyperlinks to footnotes and, since April 2006, hyperlinked cross-references to other articles and to notices about articles not yet translated or not yet assigned for translation. The full text is searchable, with word stemming[5] and character mapping[6] and non sensitivity to case, and certain metadata fields for each article (English title, French title, translator's name, or category of knowledge) are also searchable. Display of inline images (only in two articles so far) was implemented on November 28, 2006. SPO has

---

[2] http://spo.umdl.umich.edu/

[3] http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/

[4] This is very close to the Creative Commons Attribution-NonCommercial-NoDerivs 2.5 license (http://creativecommons.org/licenses/by-nc-nd/2.5/), except that the Diderot license requires users to be non-commercial, whereas the Creative Commons license requires the use to be non-commercial.

[5] For example, a search for "encycloped*" will find "encyclopedia" and "encyclopedic".

[6] For example, a search for "encyclopedie" will find "encyclopédie".

investigated providing links to images of plates in the original Encyclopedia, but work on this has been postponed repeatedly due to the complexity of the problem and of implementation.

After receiving the monthly batch of translations from Jen, SPO staff verify the metadata in the Header Database and the formatting of the text. The articles are saved in RTF format with *spostyles* applied and converted using the *makepub* script, and cross-references in previously published articles are edited to reflect the presence of new translated articles. In addition, SPO also makes occasional corrections to previously published articles based on emails sent by the Project Directors and users of the site.

It is estimated that approximately 10 hours per month of SPO staff time are spent adding new content to the Project, plus approximately 10 hours per year is spent on implementing new features. In addition, Jen spends approximately __ hours and Dena __ hours on screening Volunteers, tracking assignments, and editing translations.

Shortcomings of the current system

Scalability

SPO currently receives approximately 10 articles per month to publish on the Project website. The number of Volunteers for the Project is likely to increase over time as scholars become more comfortable with the Internet, so we can expect the number of translated articles also will increase. The number of staff hours spent updating the Project website is indirectly related to the number of translated articles submitted: each article requires additional time, but there is a fixed labor cost regardless of the number of articles in a batch. Because the Project website is updated monthly, rather than quarterly or even less frequently, as other SPO serial publications, it demands more time per year than others.

Technical deficiencies

DLXS has deficiencies when used as a platform for dynamic publications, as opposed to searching and browsing of static publications (usually digitized from print), as has been its primary use.

DLXS was not designed for content that is revised by authors or editors after initial publication. SPO staff must make these revisions by hand, and they do not show up on the Project website until the next monthly update occurs. In addition, there is no mechanism to track these revisions to content. While this deficiency is shared among all SPO publications, it is particularly acute for the Project, which sees frequent revisions to its content.

DLXS also has no built-in mechanism for linking between articles. A mechanism for handling cross-references was added to the Project in April 2006, with four link types:

a) to another translated article
b) to a notice that an article has been assigned but not yet translated
c) to a notice that an article has not yet been assigned for translation
d) to a notice that an article was cross-referenced in the original *Encyclopédie* but did not, in fact, exist in the original.

However, due to the architecture of DLXS, there is no way automatically to turn (b) into (a) once the article is published or (c) into (b) once the article is assigned. These changes must be made by hand in the electronic text, which is especially disappointing since the assigned ID number cannot be recorded in the electronic text until the article is translated and published, so

known information cannot be immediately recorded.[7] Because of the manual tracking of article numbers, this implementation of cross-references will become increasingly unmanageable as the Project grows in size, possibly forcing the Project eventually to adapt a model like ARTFL, where cross-references perform a search on the headword rather than attempting to link to a particular article.

Model of collaboration

Since Jen generally assigns articles for translation before translation begins, a Volunteer commits to translating the whole article by himself or herself unless he or she can find a co-translator to help. While this notion of discrete authorship is useful for scholars preparing CVs, in which clear evidence of contribution is generally expected, it does not allow the kind of small-scale collaboration evidenced in Wikipedia, whereby contributors improve each other's work, often one phrase at a time. The Project Directors freely edit the text before and after initial publication, but Volunteers are not able to see evidence of these changes.

Desire for new features

The Project Directors, contributors to the project, and even SPO staff have expressed interest in adding new features to the site. Among them:

a) Allow browsing of all categories of knowledge from the normalized database created by Ben Heller in 2004.
b) Release of content under a Creative Commons or other non-exclusive license for copyrighted content.
c) Allowing browsing by category of knowledge directly from hyperlinks in article headers.
d) Including links to biographical informational on each author of the original French articles (if the Voltaire Foundation can be convinced to digitize *The Encyclopedists as a Group*[8]).
e) Linking to ARTFL's images of plates from the original *Encyclopédie*, or addition of these to the Project as separate translations.
f) Addition of citations for the sources on which the original French articles were based.
g) Addition of citations for scholarly commentary on particular articles in the *Encyclopédie*.
h) Addition of sample syllabi and other materials for use in teaching the *Encyclopédie*, or citations to this material.

Feature (a) is hosted on the ARTFL site, so a link to this could be added to the current Project website. Feature (b) has been planned for all SPO publications but still not implemented, partly because it requires revisiting all content licenses. While (c) could be implemented now with little difficulty, and (d) could also if the Voltaire Foundation can be convinced to digitize its biography, (e) will take a large amount of planning for the technical implementation and to translate enough text to allow access to these images for those who do not speak French.

---

[7] While it would be convenient to include the cross-reference's ID as soon as it's assigned, there needs to be a way to distinguish bogus IDs from references to still -untranslated articles. Therefore, they can only be included in the referring article once the article referred to has been published.

[8] Kafker, Frank A. The Encyclopedists as A Group: A Collective Biography of The Authors of The Encyclopédie (Oxford: Voltaire Foundation, 1996). Pp. xxvii + 222. £50.00.

Features (f) and (g) are especially suited to contributions by users of the site other than the original translators of a given article, such as that made easy by wiki technology. The Project Directors have recently brought up feature (h), and SPO has suggested setting up a CTools site for these resources, which are not easily hosted in the DLXS framework.

Proposal to use wiki technology

Background

A wiki "is a type of Web site that allows the visitors themselves to easily add, remove, and otherwise edit and change some available content, sometimes without the need for registration. The ease of interaction and operation makes a wiki an effective tool for collaborative authoring."[9] The most famous wiki is Wikipedia,[10] though not all wikis need operate under such an open model for collaboration.

Wiki technology makes it easy for Volunteers to improve the work of others, and provides a simpler model for cross-references than the current system that does not require assigning unique IDs in advance.

In earlier discussions about using wiki technology, the Project Directors expressed strong reservations about making the content of the Project publicly editable. While the most famous wikis are open for editing by all, the Project's wiki need not be.

In addition, by using a wiki, Jen would no longer need to track licenses from Volunteers since the wiki software could be configured to require users to agree to license their content. But best of all, wikis allow for user-driven innovation of the structure of the site, which would allow Volunteers to begin adding new features of their own through the site.

Choice of wiki software

Any wiki software used must be open-source and depend only on other open-source software. Wikipedia, for example, uses MediaWiki,[11] and U-M Information Technology Central Services (ITCS)[12] offers MediaWiki as a web application integrated into U-M's LDAP infrastructure.[13] The software chosen must be flexible in the terms under which contributions are licensed to users so that an appropriate license can be chosen by SPO staff and the Project Directors.

Since each article translated has its own metadata (French and English titles, category of knowledge, and translator information) that can be searched and browsed in DLXS, it would be desirable to maintain this fielded data, with controlled vocabularies when applicable, in the wiki. "Structured wikis"[14] promise to offer such capability. Some options to investigate are:

- WikiD[15]
- ThingDB[16]

---

[9] Wiki. (2006, November 22). In *Wikipedia, The Free Encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Wiki&oldid=89448131

[10] http://www.wikipedia.org/

[11] http://meta.wikimedia.org/wiki/MediaWiki

[12] http://www.itcs.umich.edu/

[13] http://um.web.itd.umich.edu/index.php/Wiki

[14] Structured wiki. (2007, October 29). In *Wikipedia, The Free Encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Special:Cite&page=Structured_wiki&id=166929898

[15] http://www.oclc.org/research/projects/wikid/

[16] http://demo.openlibrary.org/about/tech

- TWiki[17]
- MediaWiki *templates*[18] and *categories*[19]

Technical specifications

SPO envisions a wiki for the Project that is publicly readable but editable only by Volunteers, the Project Directors, and SPO staff. The Project Directors would continue to screen Volunteers, adding them to a list of users authorized to edit. Edits would be tracked and recorded for each user, as with all wikis.

Each article in the Project would constitute a *page* in the wiki. All users adding or editing content on the site would be prompted at the point of creation or editing of a page to agree to the terms of a license for redistribution of the content.

Articles could be monitored for changes using the *recent changes* feature[20] of MediaWiki or similar feature in other wiki software. Others besides the Project Directors could share the burden of verifying changes. In addition, any contributor could choose to watch pages of interest—most likely articles they originally translated but possibly also others.

There are four types of cross-reference links in the current system:

a) to another translated article
b) to a notice that an article has been assigned but not yet translated
c) to a notice that an article has not yet been assigned for translation
d) to a notice that an article was cross-referenced in the original *Encyclopédie* but did not, in fact, exist in the original.

These can either be replicated in the wiki, maintaining the system of assigning articles for translation ahead of time, or the model could be adjusted somewhat to fit a more common wiki model. Such a system would allow any Volunteer to add any new translation on demand, conflating (b) and (c). While a Volunteer's planned translation might get "scooped" if articles cannot be "reserved" by a Volunteer for translation, having it this way might also encourage Volunteers to complete more quickly the translations they have begun. For type (a), volunteers would help maintain *redirects* ("see" references) ensuring that when the title of an article is translated differently in another article, the link will still go to the right place, and also allowing variants of a title to have equal prominence in browse listings. For type (d), one brief page would be created for this circumstance, and users of the site could even list articles linking to it (that is, list articles with references to non-existent articles).

Besides having "dead-end" cross-references, the *Encyclopédie* also contains references to headwords for which there is more than one article and possibly subarticles, with unclear boundaries between them. Kevin led an effort in 2004[21] to split translations into the smallest chunks identified by ARTFL so there would always be a one-to-one correspondence, but in the

---

[17] http://twiki.org/

[18] Help:Template. (2007, October 29). In *Meta-Wiki*. Retrieved from http://meta.wikimedia.org/w/index.php?title=Help:Template&oldid=711066

[19] Help:Category. (2007, October 29). In *Meta-Wiki*. Retrieved from http://meta.wikimedia.org/w/index.php?title=Help:Category&oldid=721348

[20] Help:Recent changes. (2007, October 29). In *Meta-Wiki*. Retrieved from http://meta.wikimedia.org/w/index.php?title=Help:Recent_changes&oldid=640735

[21] Kevin worked with Ben Heller to implement a new numbering scheme, with some articles split into smaller chunks to correspond to ARTFL chunks, resulting in retired ID numbers with no content. Since the Project did not yet offer permanent URLs for the content, only a handful of incoming webpage links made by hand were broken in the process.

current system this leads to multi-part cross references, such as "*See* Faith [1][2][3][4][5][6]". MediaWiki handles these easily by allowing *subpages*[22] that automatically link to the parent page; other wiki software probably has a similar feature. This is better than links to *sections* in MediaWiki pages, which work like anchors in HTML and therefore break easily.[23]

In case a structured wiki will not allow for the desired fielded data in the wiki, metadata elements in the current Project could be handled using other means:

Metadata element	***MediaWiki implementation***
category of knowledge	could use MediaWiki categories, but tying these to the full controlled vocabulary would be difficult
the author of the French original	could be conveyed in the body of the wiki page using MediaWiki templates
the URL of the ARTFL French text	could be conveyed in the body of the wiki page using MediaWiki template
Volunteer's name	would not be displayed with the wiki page, but username would be retrievable by viewing the page's edit history
Volunteer's affiliation	would not be displayed with the wiki page, but the user could record this information on his or her *user page*
Volunteer's email address	would not be displayed with the wiki page, but the user could record this information on his or her *user page*

The Project contains a handful of articles translated not by contributors to the Project but by authors of previously published print books, for which permission was obtained to reprint these texts in the Project. In theory these articles have not been edited since then by the Project Directors or by SPO staff making edits, though inevitably some changes have been made. In any case, these articles should be locked from further editing by all users except the *administrators* for the Project's wiki.

The current Project website contains permanent URLs for individual articles, and SPO does not wish to stop supporting these URLs. They can be redefined to point to the wiki URLs; however, to avoid ongoing maintenance of these permanent URLs, we should add a wiki comment to all existing articles, noting that a redirect must be set up in case the article is ever renamed so that the permanent URLs will continue to work.

Redirects and *piped links* in MediaWiki will allow an article title to be translated in more than one way by the original translator, giving each title equal priority, and also allow others to refer to that article by a different title than the one used by the article's translator.

---

[22] Subpages. (2007, November 9). In *Meta-Wiki*. Retrieved from http://meta.wikimedia.org/w/index.php?title=Subpages&oldid=608569

[23] Help:Section. (2007, November 9). In *Meta-Wiki*. Retrieved from http://meta.wikimedia.org/w/index.php?title=Help:Section&oldid=736405

Mathematical formulas, displayed in DLXS using images, could be encoded in TeX notation if MediaWiki is used.[24]

Other desired features can be implemented in the text of article pages, possibly using MediaWiki templates, and by creating wiki pages with additional content that do not include text translated from the Encyclopédie.

Challenges for implementation

There are a few important issues to consider in the implementation.

First, the sheer number of articles in the *Encyclopédie* will continue to present a problem for the Project since, for a given headword, there is often more than one article or subarticles. If Jen will no longer be involved in assigning articles, the Project will need a tutorial on disambiguating articles found in ARTFL and perhaps an interface that allows a user to supply an ARTFL URL and see if it, or any portion of it, has been translated already. After all, if two Volunteers translate the same article's title differently, the second won't realize that it was previously translated in the Project.

Second, the Project might decide to freeze the current project website and keep it online, only adding content to the new wiki site. There two reasons for this:

a) Some content rightsholders might not allow their content to be put in the wiki, but the content might be valuable enough that the Project wishes to keep it online and accessible.
b) SPO could add a notice about the new site to every article on the old site but not worry about attempting to provide a path to the equivalent article in the wiki. Doing so would skirt the question of what happens if a wiki page pointed to by a URL is changed significantly to the point where it no longer resembles the original in DLXS.

Third, MediaWiki's search system does not handle arbitrary stemming as DLXS does; however, since major search engines do not do this either, it seems that few users expect this functionality.

Fourth, if Jen will no longer assign articles for translation, we need to determine whether currently assigned articles will have a special grandfathered status in the wiki, whereby the Volunteer has a right to first translation.

Preliminary steps to implementation

Phase I: Prototype development

Kevin will hire an unpaid SI student intern during a winter term. Phase I should be as follows:

1. SI student evaluates wiki software and recommends which to use.
2. SI student installs this software on scholarlypublishing.org at a restricted URL and tests creating articles with Greek text, formulas, and images, with all metadata recorded in new format. Student demonstrates equivalent functionality of browse and search pages and all desired levels of editing permissions.
3. If time permits, student proposes a method for linking to images of plates.

[24] Help:Displaying a formula. (2007, October 29). In *Meta-Wiki*. Retrieved from http://meta.wikimedia.org/w/index.php?title=Help:Displaying_a_formula&oldid=724824

Phase II: Migration

Before moving forward, the Project Editors must agree that the wiki is superior to the current Project site in DLXS. This work could be done during a spring/summer term by the SI intern paid hourly and possibly receiving additional intern credit, or by SPO staff. Phase II should be as follows:

1. Resolve the "challenges for implementation" outlined above and any other issues raised during the prototype development.
2. If necessary, Project Editors contact rightsholders for permission to release content under a more permissive license.
3. Write and test a script to take Text Class XML files from DLXS and insert them into the wiki.
4. Freeze updating of old site.
5. Run script to migrate frozen site.
6. Add "meta" pages (information about the Project, wiki editing help) to wiki.
7. If not already completed and time permits, propose a method for linking to images of plates.

Phase III: Clean up

This work will be done by SPO staff:

1. Change permanent URLs (both name resolver and handles) to point to wiki pages (if necessary).
2. Take down DLXS collection and set up redirects.
3. Kevin takes Diderot-specific code out of makepub.sh and deletes Diderot-specific scripts from bin/s/spocolls/.

Conclusion

Despite the challenges, moving the Project to a wiki should allow the Project to grow more naturally and allow the Volunteers, Project Directors, and SPO staff to experiment with new models of scholarship. It should also save the Project Directors time spent on the Project, allowing the community of Volunteers to pick up some of the work.