Project Endings

Notes from study day, 17 May 2016

1. Overviews of the four projects on which our case studies will be based.

* Janelle’s overview of *MoEML*.

Assets:

Gazeteer; mapography of 92 maps of early modern London; Joey produced a finding aid for mortality bills; also edited texts and edition of John Stow’s survey of London marked up to high scholarly standards.

What would be lost without preservation:

17 years of work by various scholars, including emerging scholars. Resource used regularly. Widely linked from courseware systems, genealogical sites. Extreme detail in the gazeteer, a model of granularity. Exemplary documentation.

Preservation challenges:

Re-encoded a number of times. Future-proofing needed.

* Claire’s overview of *Le Mariage sous l’Ancien Régime*

(See GitHub for the long version.)

Assets

Annotated texts: persons and places are linked to an index containing hundreds of entries. Numbered notes provide commentary on ambiguous, little known or difficult concepts. Commentary in the form of longer articles also links to the texts and images. An extensive bibliography. The 70 images are also annotated and searchable thanks to the Image Markup Tool developed by Martin. The search engine will be optimized to perform searches using early modern spelling. The redesign of the interface for the static buildis well on its way.

What would be lost without preservation:

Besides years of HCMC work and and SSHRC and UVic funding, a research tool that can prove useful to Early Modernists across disciplines for years to come.

Preservation challenges

Time is needed if the project is to reach a satisfying endpoint. Proofreading texts and metadata, annotation, and the addition of links to external documents are all ongoing

Link rot is an issue given the hundreds of hyperlinks in the indices.

* Elizabeth’s overview of the *The Robert Graves Diary Pro*ject

Context

Editions present arguments about the material—this is a more of an archive.

Boutique software a problem but exist db allowed the project to move on.

Letters added to diaries; Schreibman’s work on the letters took off.

The gap between digital representation and text is important: for example, the materiality of ink, scored paper, etc.

Assets:

High res scans of every page.

Beryl Graves transcript more legible.

Enclosures that came with diaries.

Search engine.

The project appears in TEI guidelines. Finished in 2004, it still holds up.

Aps allow drilling down.

Preservation challenges:

The editorial apparatus needs to be revisited, and annotation added. Convert to TEI P5 from P4.

* Ewa’s overview of *The Nxa'amxcín Database and Dictionary*

Context

Working with the Colville tribes in Washington State. There are 3 major language groups among the Colville tribes, and 23 languages in the Salish family, 7 in the interior.

The dictionary answers the fundamental linguistic question: what does the language look like?

The first goal of the project was a print dictionary for linguists and native speakers, prepared by using 3 x 5 file cards showing roots and the words emerging from them. In 1991 there was a Lexware transcription. With the use of Word Perfect and DOS, there were potentially substantial losses of this work by the late1990s. Greg and Martin were able to retrieve this data.

In considering the electronic v. the print dictionary: usability issues haven’t been explored by linguists.

Assets:

This is the only written record of the language.

This is only lexical resource in the world that uses TEI. Other digital lexical resources use boutique approaches.

What would be lost without preservation:

A wealth of knowledge about the language, encoded in the language.

Preservation challenges:

Searchability a challenge because of fonts; too much data is returned on searches.

Should preservation include UVic servers? Password protected: no permission for open access from outside the Nxa'amxcín community. Conversations about preservation and access will be challenging.

The new dictionary uses an XML database, but not TEI as TEI is difficult for community members.

2. Introduction to technical issues by Martin

Demonstration of a website created in 1996

It still works because the HTML and JavaScript still work. Both are designed to stay, as is CSS. All backward are compatible.

JavaScript ES6: no deprecation! CSS has moved away from versions to CSS modules that proceed on their own.

Of 17 links, 11 work on Wayback Machine; only 7 on the current web. Fragility of URLs has to be noted.

Plans for static builds of our case studies

* Everything stored in subversion;
* Editors store and retrieve from there;
* Programmers add code in the same repositories;
* Editors and programmers upload into eXist-db.

### Current web apps in eXist-db

Browser to controller (fields request) to XQuery. Controller and XQuery inside db. XQuery fetches XML data, sends to XSLT for conversion … and back.

### AJAX links: “What you think in your doc is not in your doc.” Information in different places; pointers.

Problems with model:

* Forgetting to upload to eXist.
* Web aps are thus difficult to archive.
* Taking a single doc from the collection is tricky – but survival counts on others using our data. If “all the other crap” is necessary then…
* Search engine has to be crafted specially for each ap, so transferability.
* Versioning is incoherent and editions don’t exist. **Never** an up-to-date coherent version. Searches on the fly…Impression of searching a collection of coherent docs but NOT the case. Google is better as it sees the collection as a website, which it isn’t.

Static build principles

* Build everything all the time
* Validate and diagnose relentlessly
* Make every doc coherent and complete
* Duplicate
* Make HTML docs that degrade gracefully (turn off CSS, or HTML).
* Create every possible version of docs you can imagine being useful.

### PDFs another preservation strategy!

MoEML Build Process

* Validate source XML
* Create a better version (original XML)
* Add generated XML
* Validate original XML
* Create “standalone” versions of all original XML docs
* Validate standalone XML
* Create standard XML: more normative versions of XML docs for those not interested in more focused queries.
* Validate
* Create TEI simple versions (best for early modern print)
* Validate
* Create TEI lite versions
* Validate
* (Interchange: mediated re-use – rather than interoperability. Simplification = one meaning of the encoding, yet another angle.)
* Create KML output from all location files
* Validate
* Create all the fragments required for responses to AJAX requests
* Create HTML5 versions of docs.
* Validate

### What standalone HTML contains. (lots)

### Role of Jenkins: seeks changes from subversion every 5 minutes

Advantages of static build

Future users can take want they want, in a self-contained package, in the format they need. And it works without a web server – except for searches!

3. Stewart and Greg on other issues in preservation using static build

Full text index: offset/address method. E.g. doc #, line #, word #.

A dynamic search engine is important, but in static build there is no database engine. Simulacrum version of that capability without search engine infrastructure. Where are the thresholds? Every search engine is bespoke. A client-side search engine? Never as good as search in db.

Google and gaming search engines: we need the librarians to explore this question.

4. Joey on using diagnostics

Generated file displays common encoding errors not caught by validation. Validation doesn’t handle linked data.

What can validation do?

Put i.d.s in proper form

MoEML will integrate into static build.

Fuzzy errors: duplicate person entries, e.g. do two people have the same name?

5. Corey on defining digital preservation – or whatever we’re doing…

The multiplication of digital projects is a huge challenge for library repositories. All basic terms need definition, including the notions of “community” and “public.”

The challenge is less tech and tools than money for personnel to build and maintain infrastructure.

UVic has a digital preservation working group with a working web page.

Tool sets are available, for example, Archive-It, from the Internet Archive: curated collection of web-based content. Archivematica.

A spectrum? From fully accessible repositories to sealed, deep storage spaces with no access. Selection of what to preserve is a huge issue for libraries.

### Corey has been seconded (for part of his job) to COPPUL, the [Council of Prairie and Pacific University Libraries](http://www.coppul.ca/), which has lots of recommendations to their members to help build capacity.

Comments from the group

SSHRC is developing a policy that will require applicants to budget for data management plans. Project Endings can provide templates.

We should produce a white paper detailing how library and content providers can work together on DH projects from their inception. Janelle: we can provide models for scholarly legitimacy in grant applications thanks to peer review of digital content, addressing questions of ethics, and of course preservation plans.

Martin: HCMC could build a diagnostic tool for sites to test their archivability.

6. John on digital archaeology

A fascinating history of hardware and software is available in the slide deck.

7. Lisa

The Library colleagues can create a prototype for library strategy, guidance for researchers, and requirements for researchers.

Options for indexing and search: John can work on this, along with emulation and packaging too.

8. Group wrap-up

Budget:

No change in plan to spend approximately $30,000 in fiscal 2016 to support Research Assistants for MoEML and the Nxa'amxcín dictionary. Those RAs will also contribute to the over-all goals of Project Endings. Ewa will also seek a Work Study student.

Timelines:

We will (gradually) begin to use Git as a project management tool, to track progress, assign tasks, note milestones, and collect all manner of work on the project.

Documentation

Elizabeth noted the importance of documenting problems and outcomes as we progress.

We need to decide how visible these processes will be. Joey referred to MoEML best practices using ASANA; worklogs; project logs; team meeting minutes. We will make our progress constantly visible on Git? Janelle noted that MoEML’s current documentation practices have the advantage of being searchable. Janelle suggested that Tye could chase us down for monthly progress reports.

Communication plan

Elizabeth will draft a model based on the templates for the case studies[?]