Digital Curation of Text

Kevin S. Hawkins

Webinar recording to be made available at http://connect.ala.org/node/132171.

Just slides (with clickable links) are at http://www.ultraslavonic.info/talks/20120315.pdf.

About me: where I work

MPublishing
University of Michigan Library

We provide academic publishing services that are responsive to the needs of scholars, and that foster a sustainable economic model for academic publishing. We also educate and advise the U-M community on copyright and publishing matters, and we advocate for the broadest possible access to scholarly communication everywhere.

http://publishing.umich.edu/

About me: what I do

My team manages:

- Online publication of about 20 active journals (OA and non-OA)
- Online publication of a few monograph series
- Digitization of hundreds of titles per year for ACLS Humanities E-Book
- Ingest of publisher PDFs into HathiTrust

Definitions, abstractions, and a caveat

LAYING THE GROUNDWORK

What is digital curation?

“Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.”

—“What is Digital Curation?” (Digital Curation Centre)

How is this different from digital preservation?

Digital curation “recognizes that, in order for digital assets to be maintained over a long period of time, they must not only be preserved but must also be created according to high quality standards to ensure interoperability with other data and to enable re-purposing and discovery by future users beyond the original creators and users.”

—“Digital Curation and E-Publishing: Libraries Make the Connection.” (Choudhury, Furlough, and Ray)

What kinds of digital assets?

Libraries collect what others produce and become stewards of this content.

e.g., documents in an institutional repository

Libraries create content by digitizing collections.

Libraries provide guidance to users creating digital content.

Types of activity

- **Pre-emptive intervention**: Provide guidance to users creating digital content so that it will be easier to preserve and repurpose, regardless of whether it will end up under the stewardship of a library.
- **Active management**: Make stewarded resources accessible and discoverable, and support their reuse *beyond simply viewing*.

What else do you do besides *view* content?

Let’s use text as an example:

- Search the full text (or specific parts of the text)
- Analyze the text for patterns

  word frequency, linguistic analysis, authorship attribution
- Reproduce it

  create an e-book version of a work so that it reflows nicely on your e-reader
- Remix it
- Annotate it

Beyond viewing

You can’t do all of these things in a system that only lets you view content.

You can anticipate some uses (like full-text search) and try to support them, but generally users need to be able to download the raw data (or access it through an API).

But will you be able to give them something useful? Have you chosen data formats that are good not only for preservation but also for reuse?

Digital text for preservation and reuse

I plan to show you an approach to digitizing text that attempts to support both preservation and reuse.

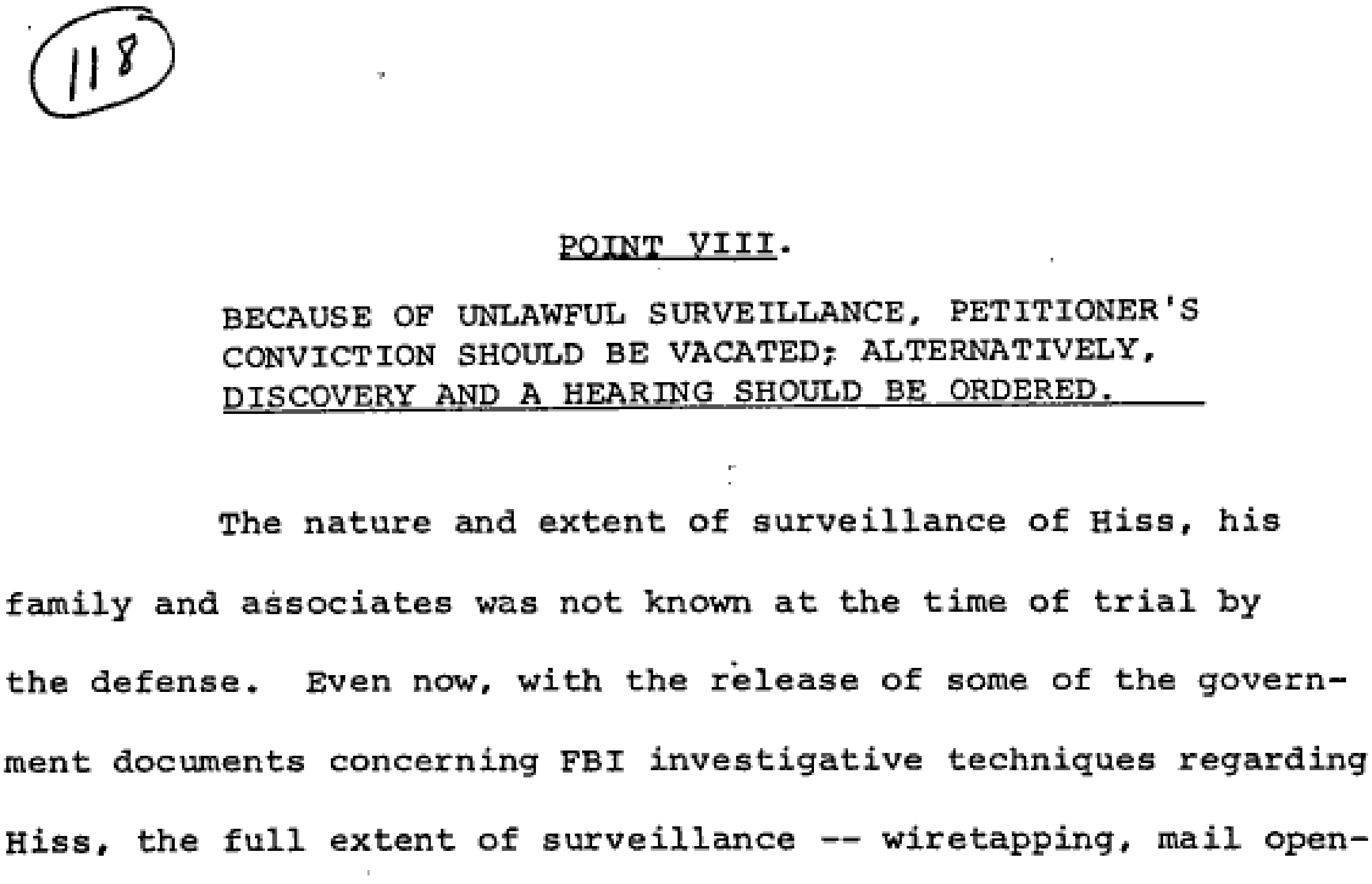
I will first discuss digital preservation techniques and then show why those alone are insufficient for supporting reuse.

Caveat

For clarity, I will assume that we are discussing text that has been digitized from a print original. The approach is equally valid for manuscripts and for born-digital documents that you wish to preserve while supporting reuse, though both of these present unique difficulties for conversion into a useful digital form.

PRESERVATION-QUALITY FORMATS THAT DON’T ENABLE REUSE

Let’s say you have an important document



118

POINT VIII.

BECAUSE OF UNLAWFUL SURVEILLANCE, PETITIONER'S CONVICTION SHOULD BE VACATED; ALTERNATIVELY, DISCOVERY AND A HEARING SHOULD BE ORDERED.

The nature and extent of surveillance of Hiss, his family and associates was not known at the time of trial by the defense. Even now, with the release of some of the government documents concerning FBI investigative techniques regarding Hiss, the full extent of surveillance -- wiretapping, mail open-

You scan it

You store an image of the page in an openly documented, non-proprietary, standard, preservation-quality format such as TIFF or JPEG2000, perhaps embedded in a PDF/A.

If you scan at a high resolution, you can print at a high quality and allow users to zoom in closely.

But the reader has to read every page. They can’t use a “find” feature to search the text.

So you OCR it

Optical character recognition (OCR) software turns an image of text into actual text that a user can search. (You’ve seen PDFs like this.)

Now you can use the “find” feature.

But what if you want to …

- Restrict your search to the author’s words, excluding words contained in quotations?
- Produce an e-book version of the document that will reflow nicely on e-book readers?
- Add annotations that are readable on any system?

We need something that is openly documented, non-proprietary, standard, and preservation-quality yet allows these things.

If you’re not happy with just full-text search

YOU NEED XML

What is XML?

The Extensible Markup Language (XML) is an open, nonproprietary format for encoding of data (or documents).

It’s not really a language: it’s a standard way of writing languages to structure data (or documents). You get to make up most of the vocabulary and syntax.

But you don’t need to invent your own! Lots of XML markup languages have been created for specific purposes. Each has a way of *validating* a document to make sure it uses the right vocabulary and syntax.

Lofts of software can process any kind of XML.

Some XML markup languages

- **XHTML**: a format for webpages
- **RSS**: allows websites to broadcast updates
- **MARCXML**: a format for bibliographic data which is more transparent to outsiders than MARC21 and which can be manipulated with standard software
- **ONIX**: a format for publishers to transmit metadata to vendors
- **JATS**: a format for publishing and archiving journal articles
- **TEI**: guidelines for representing any kind of text, especially non-digital source documents

What markup language would you use for textual documents?

If you want:

- to produce digital surrogates of a wide variety of documents
- to support uses beyond viewing
- to have the capacity to record features of the appearance of the original

your choices are:

- XHTML + CSS
- TEI

A crude comparison

XHTML + CSS

- Countless tools for creation, editing, and viewing that are very easy to use.
- Can describe many though not all features of appearance.
- *However*, has no standard way to describe structure of document beyond most basic components

TEI

- Tools are hard to use and not well documented.
- Has no standard way to describe appearance of textual features.
- *However*, has a comprehensive vocabulary for describing structure of document.

For example

XHTML

- <i>
- <em>
- <cite>

TEI

- <hi rend="italic">
- <emph>
- <title level="m">
- <title level="a">
- <title level="j">
- <foreign>
- <term>
- <mentioned>
- <soCalled>

A steep learning curve, but very powerful.

Supporting future reuse

TEI AND LIBRARIES

Use of TEI in libraries

While the TEI Guidelines offer a comprehensive vocabulary for describing text that allows scholars to encode whatever features of a text they want to represent, libraries using TEI tend to take generalist approach, encoding major structural features that everyone can agree on.

Still, encoding is expensive. Need to consider what depth of encoding is justified.

The *Best Practices for TEI in Libraries* offers a continuum between OCR and deeply encoded text.

Level 1

- Display page images to user
- Generate text through OCR
- Encoding assists with:
  - Full-text search
  - Navigation among pages

(example)

Level 2

- Display page images to user.
- Generate text through OCR.
- Navigation markers (textual divisions and headings) are encoded to allow navigation among these.

(example)

Level 3

- Generate text through OCR or conversion from digital source documents.
- Identify structural features:
  - Paragraphs
  - Block quotes
  - Footnotes and endnotes
  - Lists
  - Tables
  - Figures
- Optionally identify appearance of text: italics, bold, etc.

(example)

Level 4

- Generate text through corrected OCR, keyboarding, or conversion from digital source documents.
- Like Level 3 but with optional semantic markup within paragraphs
  - not just `<hi rend="italic">` but `<title>`, `<foreign>`, `<emph>`, etc.

(example)

By using TEI

- You are using an open, non-proprietary markup language widely used by libraries and scholars, which can be transformed to other formats (HTML, PDF, EPUB, etc.) as needed.
- In the future, you can “upgrade” the text to a higher encoding level to make it even more useful.
- A scholar can add their own markup for their own research purposes.

Further reading

1. Introduction to XML for Text
2. TEI by Example
3. Best Practices for TEI in Libraries (shows the Alger Hiss document encoded at Levels 1–4)

kevin.s.hawkins@ultraslavonic.info

@KevinSHawkins

Webinar recording to be made available at http://connect.ala.org/node/132171.

Just slides (with clickable links) are at http://www.ultraslavonic.info/talks/20120315.pdf.