

Why create TEI documents, or
when is fulltext not enough?

Let's say you have a document of literary or historical importance

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

If you scan at a high resolution, you can print at a high quality and zoom in close enough to make out things that are hard to read.

But you have to look at pages as if they were part of a physical document. You 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 of scanned documents like this.)

Now you can use the “find” feature.

The Texas Almanac for 1858

[More Sizes](#)[Lower Lights](#)

Page:

13



Search Inside

[Return to Search Results](#)[Citation](#)[Metadata](#)

HORTICULTURE AND CHRONOLOGY.

13

HORTICULTURE FOR APRIL.

Plant Cucumbers, Squash, Pumpkins, Melons, &c. of all kinds for a full crop. Plant Black-eyed and Crowder Peas, Bush, Lima and Carolina Beans. Sow Okra, Long Orange or Long Scarlet Carrot, Drumhead Cabbage for a late crop, Cabbage-head Lettuce, Radishes, Red or Yellow Top Turnip. Set out Cabbage-Plants, Lettuce, Tomatoes, Egg-plants, Peppers, and all



FULL-TEXT

CATALOG

Search words about or within the items



LOG IN

[Advanced full-text search](#) | [Search tips](#)

Full view only

[« Back to Utah State University Press collection](#)

About this Book

Alaska's daughter : an Eskimo memoir of the early Pinson, Elizabeth Bernhardt, 1912-

[View full catalog record](#)

Copyright: [Creative Commons Attribution-NonCommercial-NoDerivatives](#).



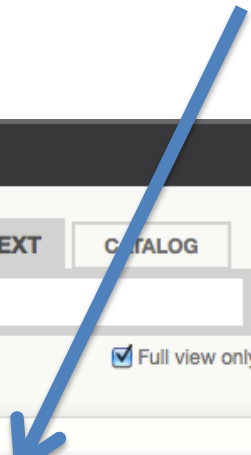
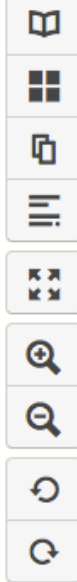
Get this Book

[Find in a library](#)

[Buy a copy](#)

[Download this page \(PDF\)](#)

Navigation controls: back, forward, search, and text input fields.



But OCR technology usually can't handle ...

- cursive handwriting
- printed handwriting
- early printed books
- poor-quality facsimiles (especially microforms)

So if you want to enable someone to search the text of such documents, you have to correct the OCR or transcribe the documents from scratch.

But what if you want to ...

- Digitize a print dictionary and allow searching on just the headwords
- Distinguish between words in the original manuscript and edits made by another hand?
- 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 your own annotations that will be readable on any system? (unlike annotating a PDF!)

To go beyond full-text searching, you need something ...

- that allows you to annotate the structure of the text
- that is an openly documented, non-proprietary standard suitable for long-term preservation and reuse of the content

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.

Lots of software can process any kind of XML.

Some XML markup languages

- **XHTML**: a format for webpages which is stricter than other flavors of HTML
- **RSS**: allows websites to broadcast updates
- **MARCXML**: a format for bibliographic data which is more transparent in structure than MARC21 and which can be manipulated with common software
- **ONIX**: a format for publishers to transmit metadata to vendors
- **JATS**: a format for publishing and archiving journal articles
- **EAD**: a format specially designed for finding aids
- **EAC-CFP**: a format specially designed for authority records for archives
- **TEI**: guidelines for representing any kind of text but especially suited to non-digital source documents

A steep learning curve, but very powerful.



For example ...

Scholarly editions

- Documenting the American South ([a sample](#))
- The Chymistry of Isaac Newton ([a sample](#))

Searching

- [The Anglo-Norman Online Hub](#)

Questions?