Topic Maps
for
Cultural Heritage
Collections

Conal Tuohy

Senior Developer

New Zealand Electronic Text Centre

www.nzetc.org

NZETC
.ORG

Website visitor statistics (daily)

- around 9k visitors
- around 70k hits
- around 30k web pages
- > 1GB traffic

Website content statistics

- 75k web pages
  - 50% represent digitised documents
    - books, magazines, letters
    - articles, chapters, sections
    - illustrations
  - The other 50% are about *things*
    - people, organisations
    - places, ships,
    - literary works
    - even a few animals!
- 3.5M hyperlinks

Resource-centric vs subject-centric systems

- "Resource-centric" systems focus narrowly on digital resources
  - a catalogue of digital items
  - everything else is peripheral or secondary

- "Subject-centric" systems can accommodate anything of interest:
  - information resources
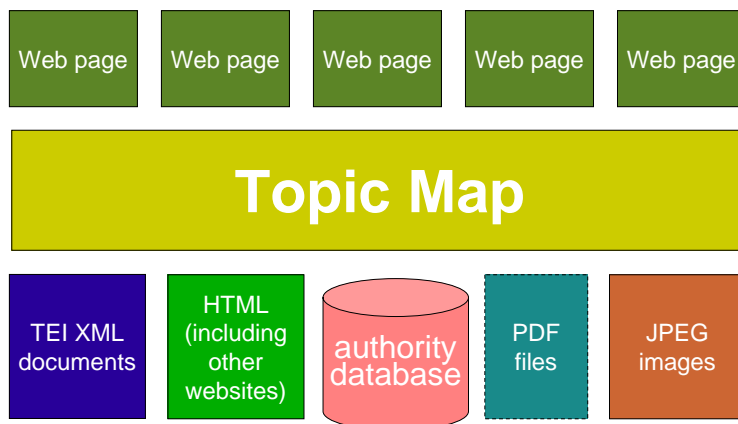 - abstract concepts,
  - or physical things
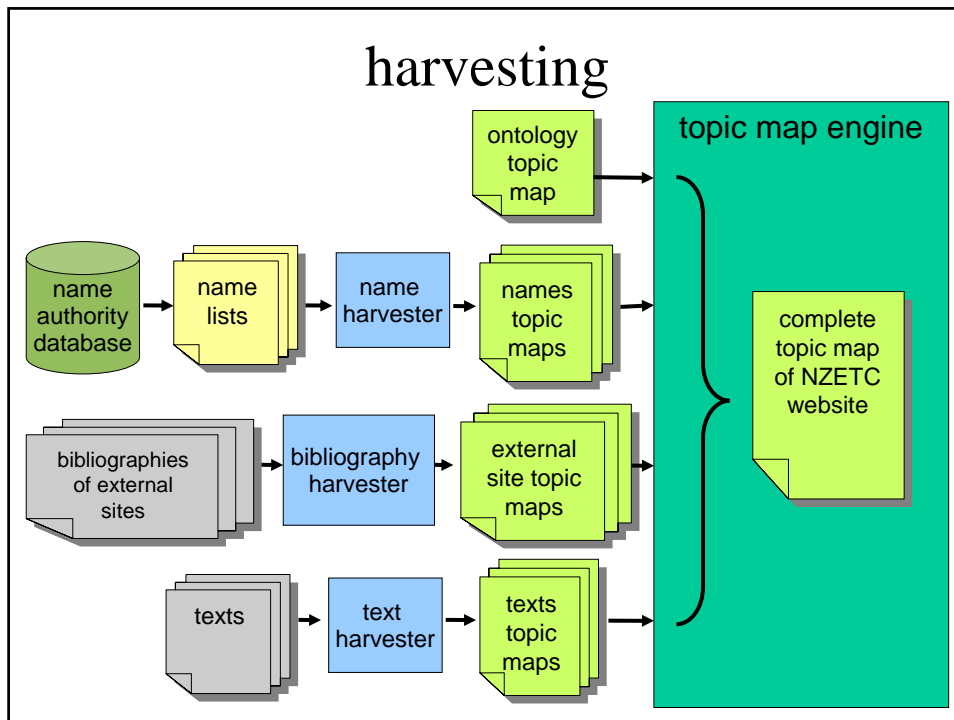
Information Architecture goals

Information resources must be transformed, chunked, and linked together into a navigable web

- Need to present information *in context* on every page
- Need an explicit model of the entire website logical structure. Not just a sitemap, but an *ontological model*
- Need to build the model automatically

so how does it work?

Topic Map layer above the digital resources

| Web page | Web page | Web page | Web page | Web page |

Topic Map

| TEI XML documents | HTML (including other websites) | authority database | PDF files | JPEG images |

harvesting

Entity Authority

- We built an authority file of entities of interest.
- We've developed a specialised database for this purpose, which we call "Entity Authority Tool Set" (EATS) to manage names and identifiers (a PSI server).
- In our digitised documents we tag every mention of these entities with their identifier. Our taggers search in EATS for a name, and select from the possible matches.
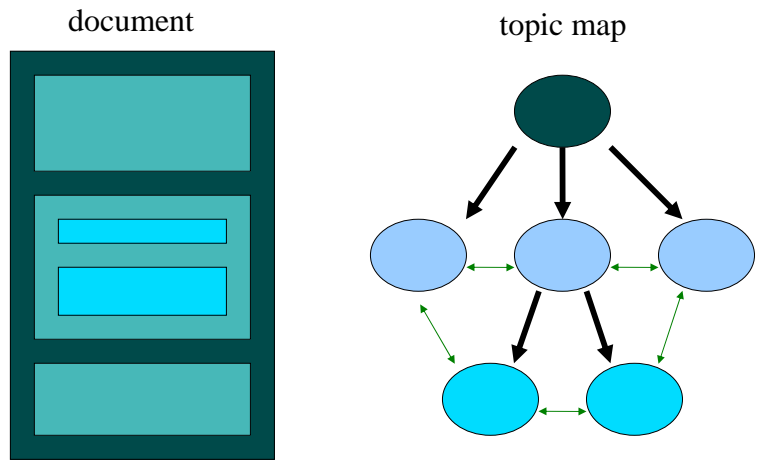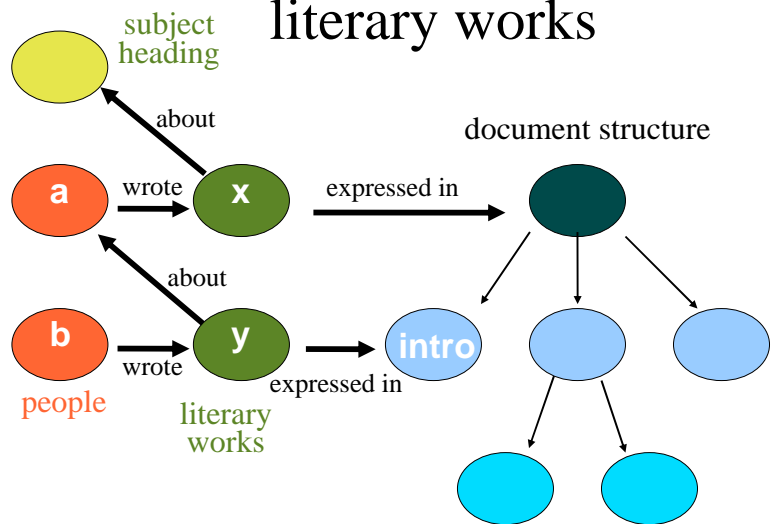
"authority" topic maps
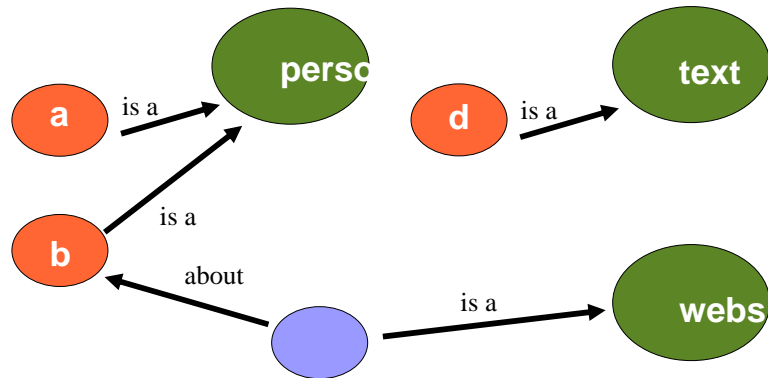
Text Encoding for Interchange (TEI)

- Bibliography
- Subject classification
- Textual structure
- Cross-references
- External references
- Commentary
- *etc.*

a document's internal structure

document

topic map

literary works

subject heading

about

a wrote x expressed in

document structure

about

b wrote y expressed in intro

people

literary works

Multiple editions of a single work

mentions, depictions, citations

Topic map statistics

- 126k topics
- 126k occurrences
- 242k associations
- 1M roles
- 115k base names
- 69k variant names (sort names)

Benefits

- Easier to provide links and contextual information
- Easy to pull together information from a variety of sources
- Implicit topics of interest are made explicit
- Improved our own understanding of our collection
- Easier to find information on the site
- Google searches work better

questions?

- only easy questions please
- contact me: conal@nzetc.org