RECYCLING OF TEXTS IN EEBO-TCP

DIGITAL HUMANITIES HACKATHON
MAY 2016 HELSINKI

EEBO-TCP

Early English Books Online (EEBO) is a collection of over 130,000 titles printed in England, Scotland, Wales, Ireland and British North America, or elsewhere in English in the period 1473-1700. The EEBO-TCP is a subset of over 25,000 TEI-compliant SGML/XML texts from the EEBO collection. These texts have been selected based on the New Cambridge Bibliography of English Literature (NCBEL). Works are eligible to be encoded if the name of their author appears in NCBEL. Anonymous works may also be selected if their titles appear in the bibliography. In addition, partner institutions have been able to request specific titles.

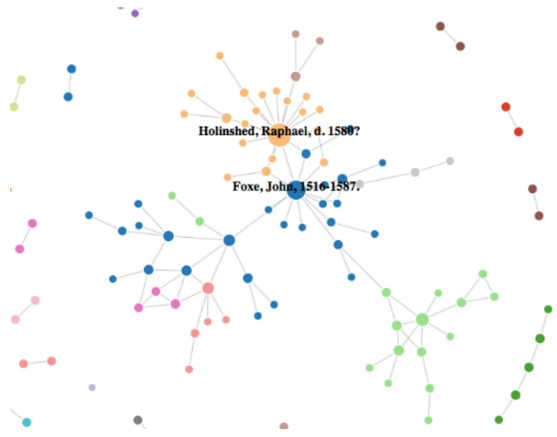
RESEARCH QUESTIONS

1) What kinds of texts were recycled?

2) Can we identify groups of texts that are "related" through their content?

Hypothesis The results of our three pilot studies should align with the findings of earlier book historical research (e.g. Barnard and McKenzie 2002; Raymond 1996; Weil and Weil 1997).

16TH CENTURY



Overlapping records: 320 (length > 1024)

Recycled text fragments show up in compilations of chronicles, voyages of discoveries, as well as biblical explications, prayers and meditations.

Various new bible translations are also prominent, as are polemical religious controversies and official documents.

METHOD

Crude normalization; same form for e.g.:

- *growe and waxe continuallie more strong …*
- *grow and wax continually more strōg …*

Identify overlaps:

- basic idea: suffix arrays
- 2.4 GB of normalized text to process
- 10 minutes of computer time

Iterative approach:

- Get non-matching single words between two matching sequences
- Deduce new rules for handling spelling variation and rerun

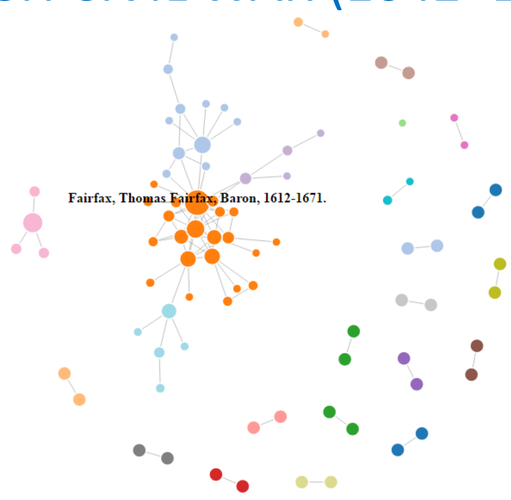
MATERIALS

Subcategory	ESTC (no of records)	EEBO-TCP (no of records)
ALL: 1473–1700	136,921	25,368
~ 16th century (1473–1700)	15,408	3,052
Civil War (1642–1651)	19,239	total 3,545 2–24 pages 1,098
Shakespeare		25,368

We selected materials for three pilot studies from EEBO TCP xml files. Comparison of the files to records in the English Short Title Catalogue (ESTC) gives an indication of the representativeness of our material.

We placed all texts on a map with metatextual information:

ENGLISH CIVIL WAR (1642–1651)



Overlapping records: 214 (length > 64)

Recycled text fragments show up in many different kinds of documents: legal, letters, speeches, religious tracts, news pamphlets etc.

Most frequent writers: parliamentarians, but also some royalists, incl. King Charles I. Most frequent texts by military and political leaders.

Royalists and Parliamentarians sometimes used the same text fragments in their pamphlets, framing them differently for their own ends.

SUMMARY OF RESULTS

The hypothesis was confirmed. All three pilot projects were able to identify the expected kinds of recycling: reprints, collections, polemical quotes, religious texts and adaptations. Thus, the method worked well for identifying stretches of recycled texts. However, visualization techniques were not able to catch all of the instances.

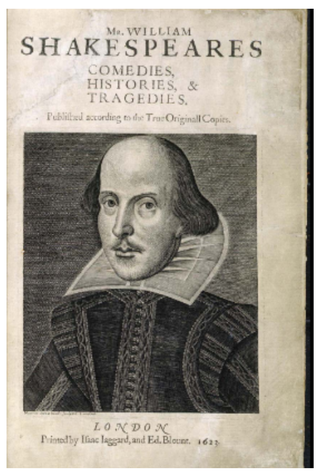
IDENTIFYING SHAKESPEARE

Which of Shakespeare's 37 plays and 154 sonnets appear in early modern English books?

Overlapping records: 56 (length > 128)

16 plays and 4 poems, Henry IV the most frequent (10 hits): demonstrates the play's popularity (Weir & Weir 2997)

Gap in 1640-62: explained by the Interregnum period (1649-60) when theatres were closed.



REFERENCES

Barnard, John and D.F. McKenzie (eds.) 2002. *The Cambridge History of the Book in Britain.* Vol. 4: 1557-1695. Cambridge: Cambridge University Press.

The Unabridged Acts and Monuments Online or TAMO. 2011. Sheffield: HRI Online Publications. Available from: http//www.johnfoxe.org [Accessed: 19 May 2016].

Raymond, Joad. 1996. *The Invention of the Newspaper: English Newsbooks 1641-1649.* Oxford: Oxford University Press.

Weil, Herbert and Judith Weil, eds. 1997. *The First Part of King Henry IV* (New Cambridge Shakespeare). Cambridge: CUP

FUTURE RESEARCH IDEAS

Results from automatic identification of recycled texts are promising.

Fine-tuning the spelling variation will give even more, and more reliable, results. Future avenues of research could focus on a larger data set which may reveal less prominent, but equally interesting, texts and authors. The focus could also be on selected topical themes, or on tracing the presence of prominent texts and authors in Early Modern English books.

UNIVERSITY OF HELSINKI
FACULTY OF ARTS

Patrik Aaltonen
Susan Huotari
Hege Roivainen
Anni Sairio
Carla Suhr
Sanna Tiirikainen

Aalto-yliopisto

Jukka Suomela
Han Xiao

1640 THE NATIONAL LIBRARY OF FINLAND

Mika Koistinen