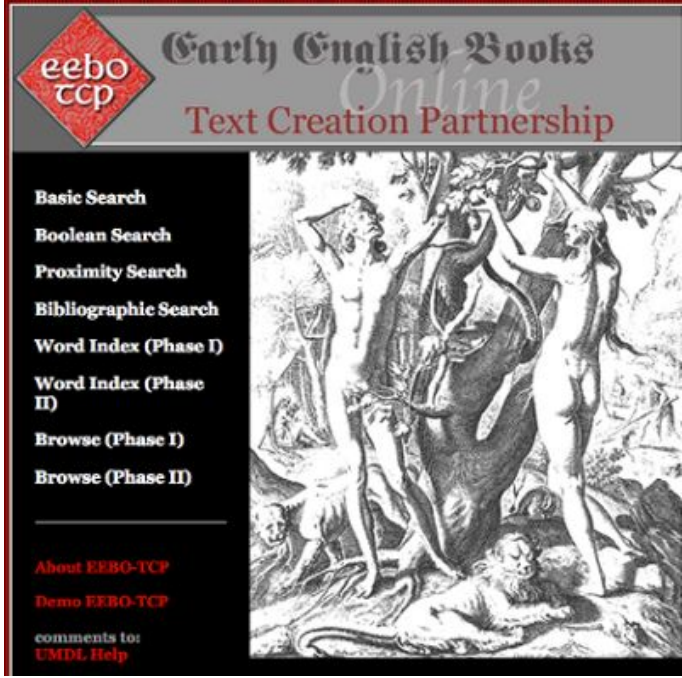
Recycling of texts in early English books



Patrik Aaltonen · Susan Huotari · Mika Koistinen
Hege Roivainen · Anni Sairio · Carla Suhr
Jukka Suomela · Sanna Tiirikainen · Han Xiao

EEBO TCP

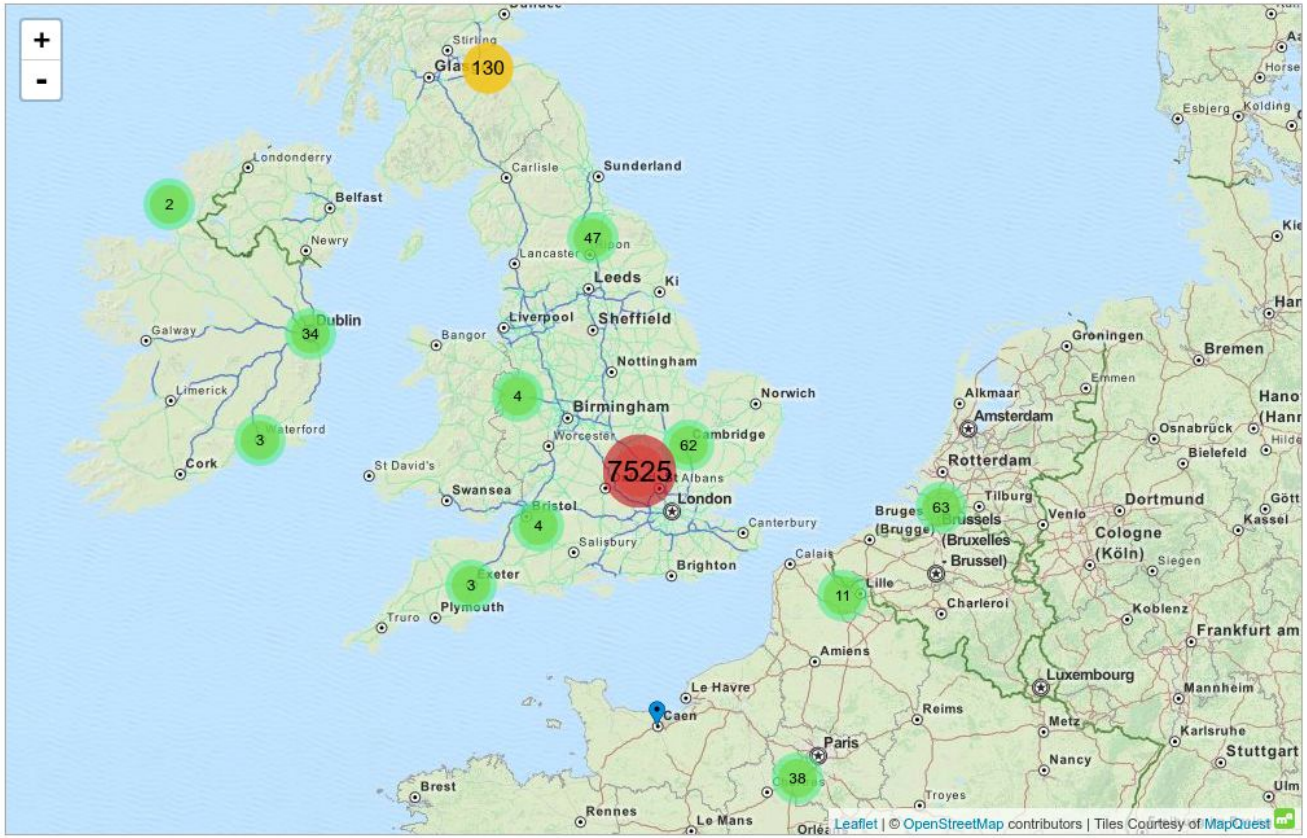
Early English Books Online (EEBO): collection of ≈ **130,000 titles** printed in England, Scotland, Wales, Ireland and British North America, or elsewhere in English in the period **1473–1700**

EEBO-TCP: subset of ≈ **25,000 titles** from EEBO

- freely available online
- 7 gigabytes of XML files, one file per title
- selected mostly based on the **New Cambridge Bibliography of English Literature**

EEBO publishers

Time: **1632 - 1661**



Research question: recycling texts

Data-driven research questions about recycling of early modern English texts:

1) What **kinds of texts** were recycled and **why**?

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

Three pilot studies:

1) Texts published in the **sixteenth century**: 3,052 files

2) Texts published during the **Civil War 1642–1651**: 1,098 files with 2–24 pages

3) Texts by **Shakespeare**

Automatically identifying recycling

Challenges: XML markup, whitespace, punctuation, **spelling variation**, typesetting errors, OCR errors, typos in EEBO TCP…

… fedde spirituallie vpon Christ, so now they féede corporallie also vpon the sacramentall bread … growe and waxe continuallie more strong in Christ … Catholike Church …

… fed spiritually vpon Christ, so now they feed corporally also vpon the sacramental bread … grow and wax continually more strōg in Christ … Catholick Church …

Identifying overlapping regions

Starting with a crude unification:

- “Catholike Church” = **KATHULIKHURKH** = “Catholick Church”

Identify overlaps in the unified text:

- 2.4 gigabytes of unified text — need to find all overlap in this material
- basic idea: construct **suffix arrays**
- ≈ 10 minutes of computer time

Iterative approach: identify what could not be matched,
develop **better normalisation rules**, repeat

Process

Three mini-groups related to the three pilot studies:

- 16th century
- Civil War
- Shakespeare

Method development in parallel:

- metadata extraction & maps
- networks analysis & topic modelling
- data normalisation …

16th Century

Henry VIII (1509–1547)

- 1534 Act of Supremacy

Mary I (1553–1558)

- devout Catholic

Elizabeth I (1558–1603)

- continuation of the protestant reformation

Age of Discoveries

16th Century

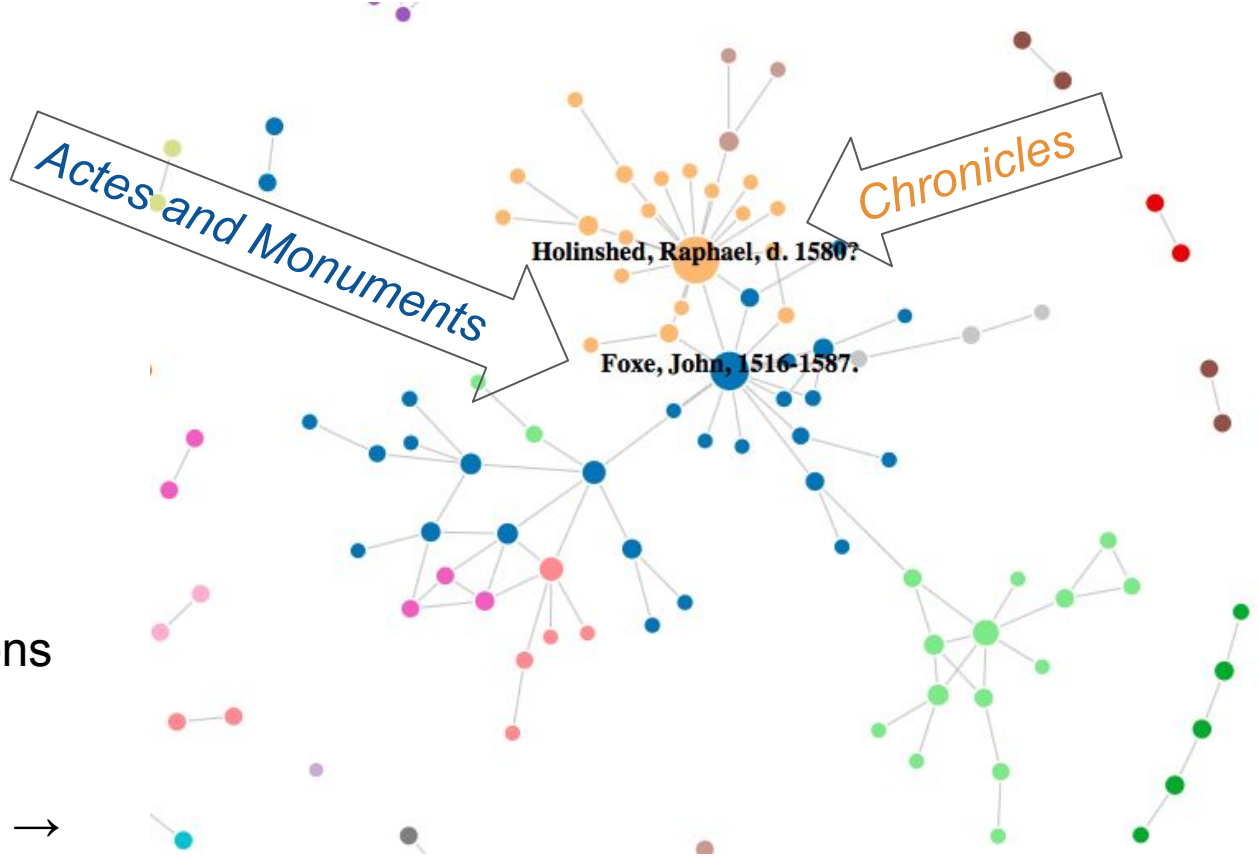
Main clusters:

- Holinshed
- Foxe

Kinds of texts:

- chronicles
- collections
- religious texts
- new bible translations
- reprints

Missing cluster: Hakluyt → “anonymous”



Text recycling during the English Civil War (1642–1651)

> “The civil wars of the 1640s were the most heavily reported conflict the British peoples had yet undergone. What historians have termed ‘the print explosion’ from 1641 played a critical part in circulating information – and mis-information – to a public thirsty for news.” (Hopper 2013, p. 15)

- The print explosion: a time of printing information and circulating it around.
- Both parties in the Civil War, the Parliamentarians and the Royalists, printed their own newsbooks and pamphlets, and both tried to manipulate the public opinion.
- An interesting time period to study!

Text recycling during the English Civil War

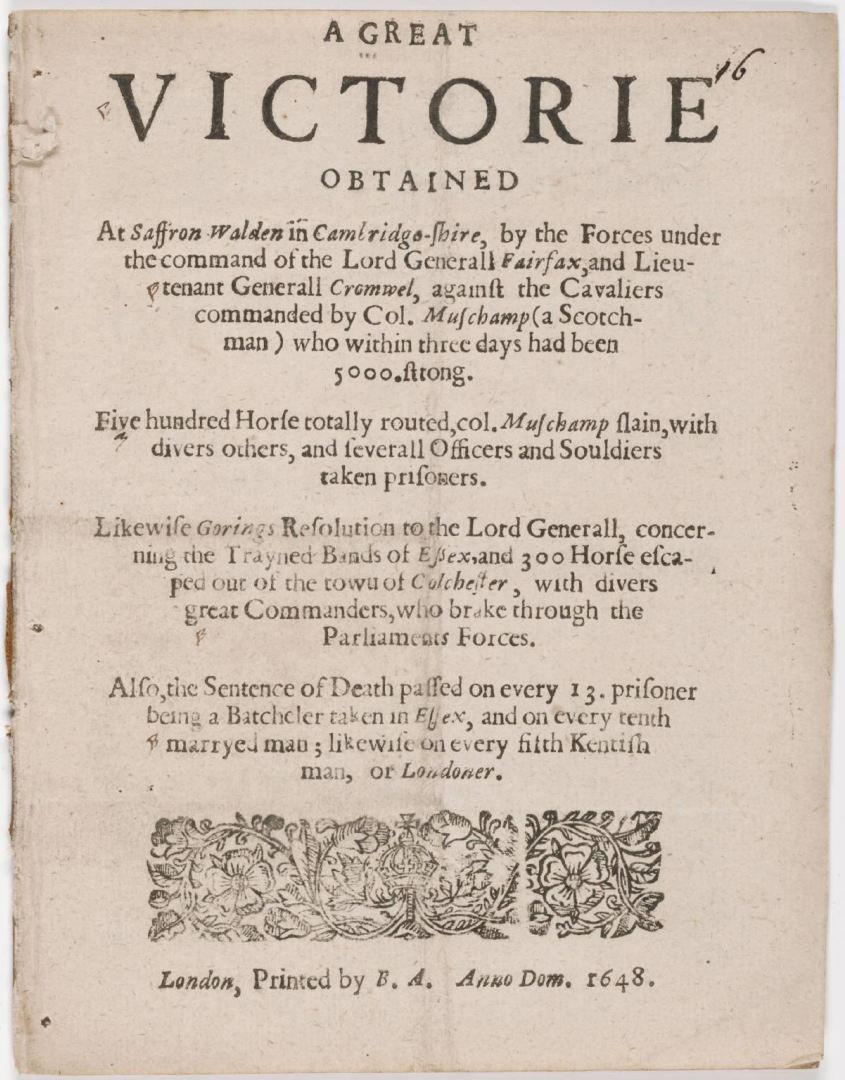
Questions:

1. What kind of documents contain recycled parts? Why were texts recycled?
2. Can we find examples of news recycling – i.e. the same news texts being printed several times?

Text recycling during the English Civil War

What kinds of documents contain recycled parts?

- legal documents
- printed letters and speeches
- religious tracts
- news pamphlets…



A GREAT

VICTORIE

OBTAINED

At *Saffron Walden* in *Cambridge-shire*, by the Forces under the command of the Lord Generall *Fairfax*, and Lieutenant Generall *Cromwel*, against the Cavaliers commanded by Col. *Muschamp* (a Scotchman) who within three days had been 5000. strong.

Five hundred Horse totally routed, col. *Muschamp* slain, with divers others, and severall Officers and Souldiers taken prisoners.

Likewise *Gorings* Resolution to the Lord Generall, concerning the Trayned Bands of *Essex*, and 300 Horse escaped out of the town of *Colchester*, with divers great Commanders, who brake through the Parliaments Forces.

Also, the Sentence of Death passed on every 13. prisoner being a Batchelor taken in *Essex*, and on every tenth marryed man; likewise on every fifth Kentish man, or *Londoner*.

London, Printed by *B. A. Anno Dom.* 1648.

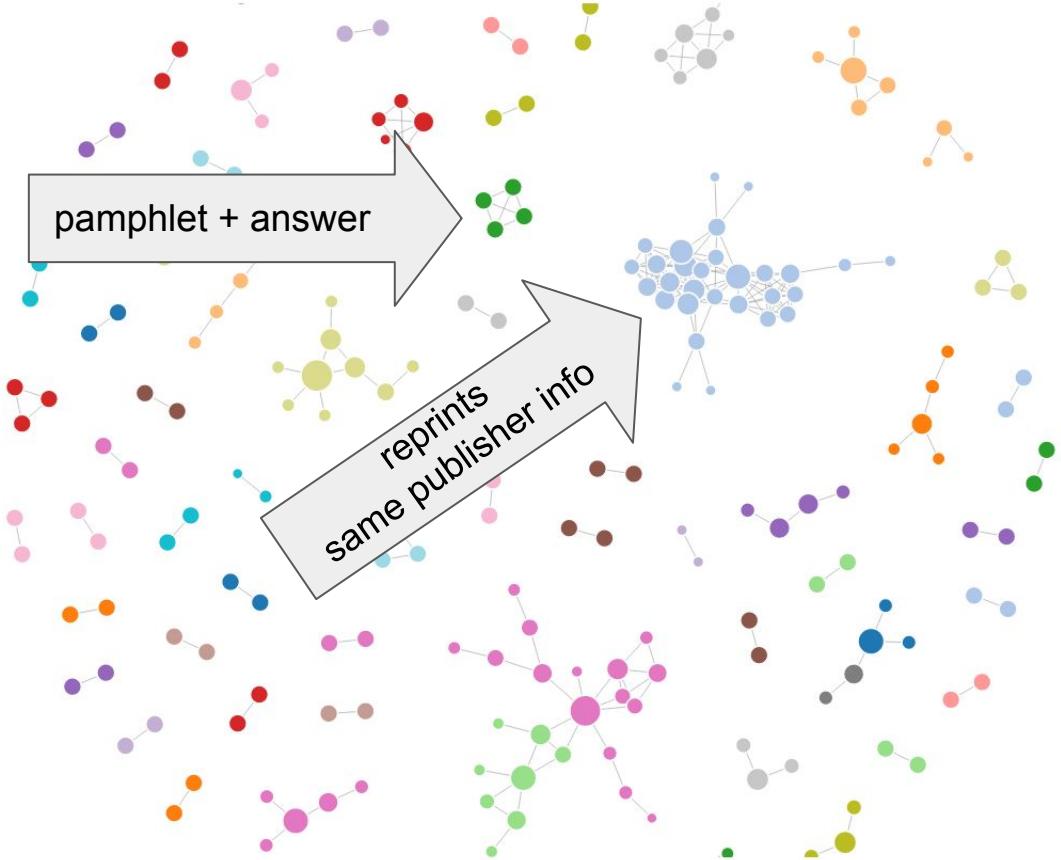
Text recycling during the English Civil War

Why are the same text fragments found in different documents? Several reasons.

- **Reprints**.
- **Text compilations**. The same declarations, speeches, letters and news reports are printed several times, in several different text compilations.
- **Quoting**. Pamphlet writers quoting and arguing against other pamphlet writers.
- **Same publisher / printer information** (short fragments).

Connected texts: a visualization

- clusters of overlapping publications



Link: https://www.cs.helsinki.fi/u/hxiao/eebo/civil-war-text/

Text recycling during the English Civil War: news recycling

From a news pamphlet printed on October 17, 1642:

On the 9. day of August we arrived be-fore Galloway, which is the strongest towne they have, except Limbrick, and there laid siege to it: so the Lord of of Clenrikard came downe and confer-red with our Lord Forbes, and the Mer-chants of the towne: they strongly re-plied, and said, that they were the Kings loyall subjects, and had not offended in the least thing, but that the souldiers in the Kings Fort had done them wrong [...].

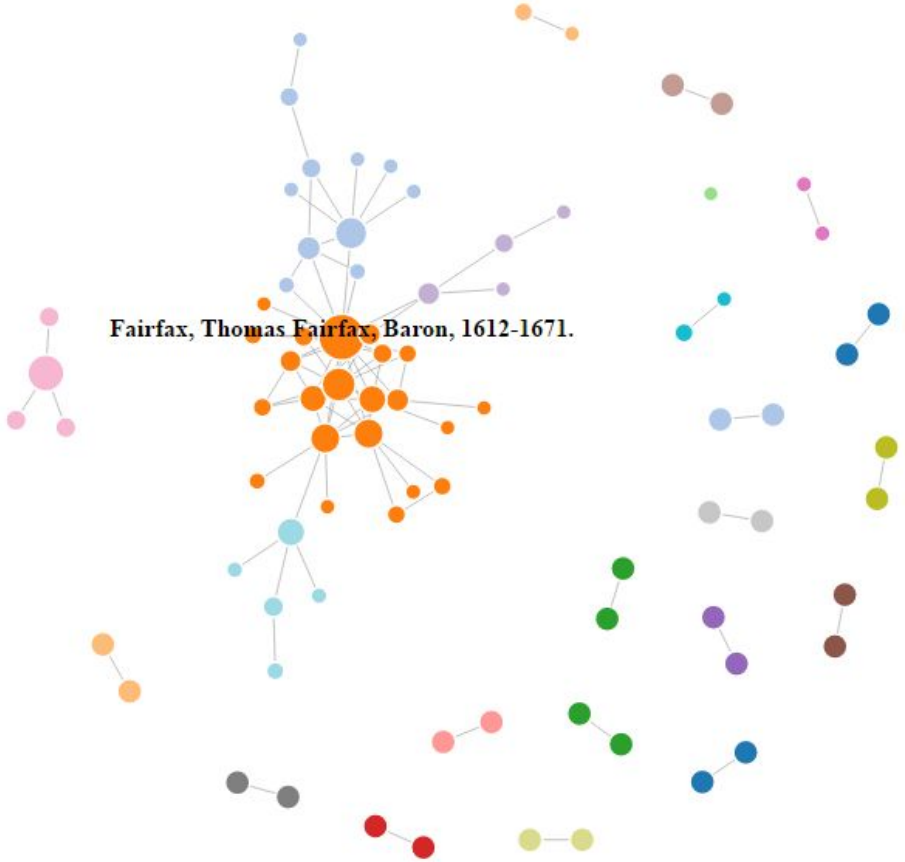
From a news pamphlet printed on October 19, 1642:

The English Fleet lately lying be-fore Galloway, which is the strongest towne they have, except Limbrick, and there laid siege to it: so the Lord of of Clenrikard came downe and confer-red with our Lord Forbes, and the Mer-chants of the towne: they strongly re-plied, and said, that they were the Kings loyall subjects, and had not offended in the least thing, but that the souldiers in the Kings Fort had done them wrong [...].

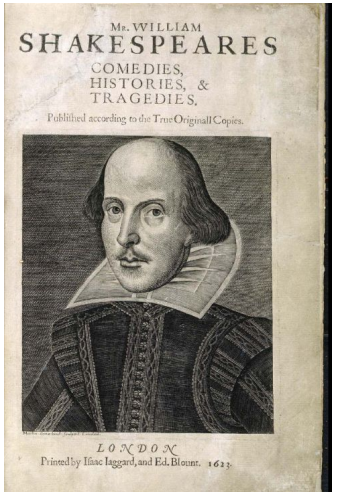
The English Civil War Author network

- 4 important clusters connected to each other
- Mainly parliamentarians
  - but also some royalists, incl. King Charles I
- Often military and political leaders
  - Thomas Fairfax, commander-in-chief
  - Oliver Cromwell, commander
  - William Prynne, political writer

Link: https://www.cs.helsinki.fi/u/hxiao/eebo/civil-war/



Identifying Shakespeare



Body of work: 38 plays and 154 sonnets
Which texts, in which books? Function? Time frame?

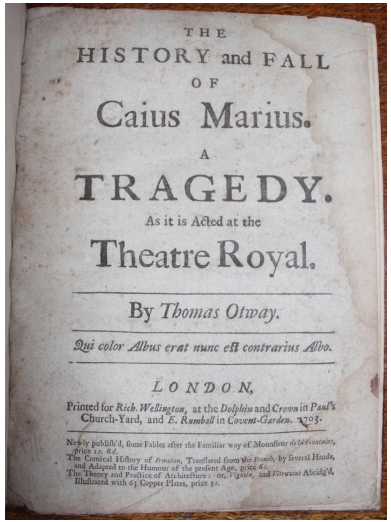
56 hits: not much!

- from 16 plays and 4 poems
- #1 is history play Henry IV (10 hits): one of WS's most popular plays in the period (Weil & Weil 1997)
- First folio (1623) the most common source
- reprints, criticism, adaptations, anthologies
- The recyclers: dramatists and theatre managers, poets, literary critics, chroniclers, hack writers

Shakespeare as inspiration:

Lavin. Oh! bid me leap (rather then go to Sylla) From off the Battlements of any Tow'r, Or walk in Thievish ways, or bid me lurk Where Serpents are: chain me with roaring Bears; Or hide me nightly in a Charnell-house O're-cover'd quite with Dead mens rattling Bones, With reeky Shanks, and yellow chapless Sculls: Or bid me go into a new-made Grave, And hide me with a Dead man in his Shrowd:

Lavinia in Thomas Otway's *The history and fall of Caius Marius a tragedy,* 1680



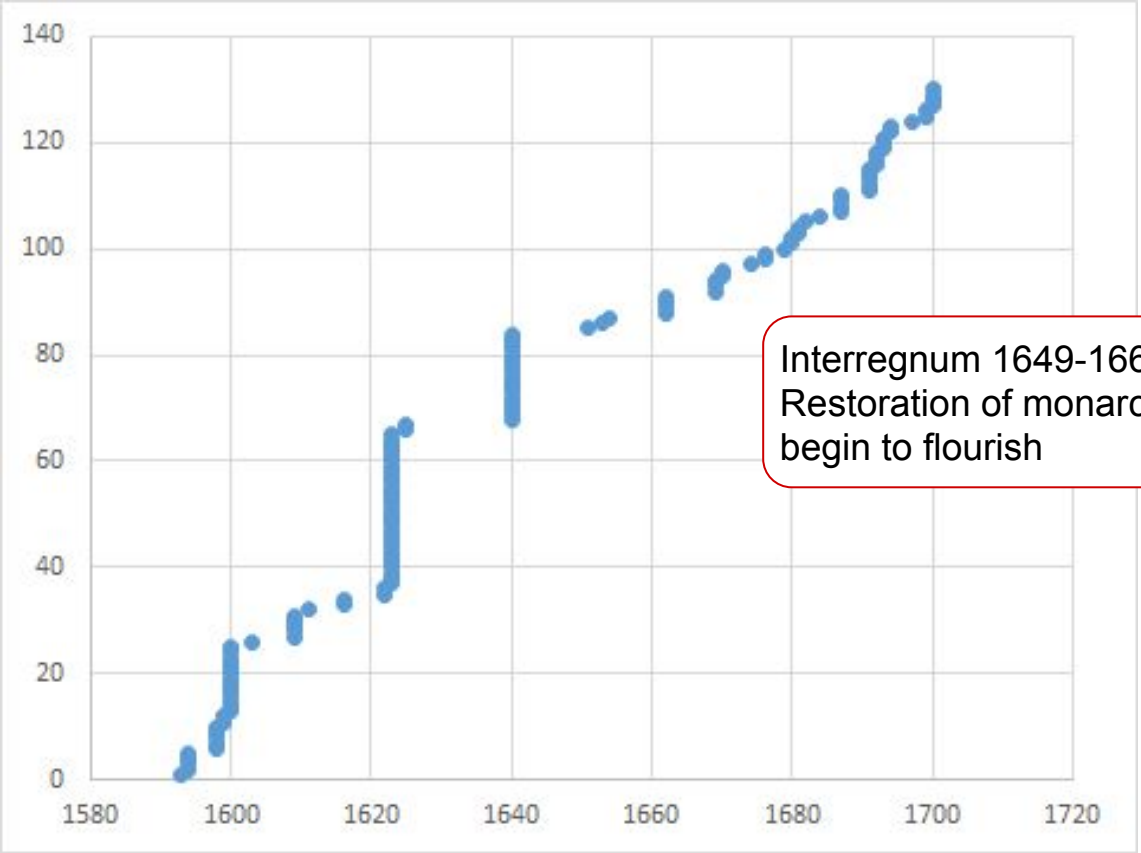
THE
HISTORY and FALL
OF
Caius Marius.
A
TRAGEDY.
As it is Acted at the
Theatre Royal.
By *Thomas Otway.*
Qui color Albus erat nunc eſt contrarius Albo.
LONDON,
Printed for *Rich. Wellington,* at the *Dolphin* and *Crown* in *Paul's* Church-Yard, and *E. Rumball* in *Covent-Garden.* 1703.

Newly publiſh'd, ſome Fables after the Familiar way of Monſieur *de la Fontaine,* price 1s. 6d.
The Comical Hiſtory of *Romans,* Tranſlated from the *French,* by ſeveral Hands, and Adapted to the Humour of the preſent Age, price 6s.
The Theory and Practice of Architecture: or, *Vignola,* and *Palladio* Abridg'd, Illuſtrated with 63 Copper Plates, price 5s.

Iul. Oh bid me leape, rather then marrie Paris, From of the Battlements of any Tower, Or walke in theeuiſh waies, or bid me lurke Where Serpents are: chaine me with roaring Beares Or hide me nightly in a Charnell houſe, Orecouered quite with dead mens ratling bones, With reckie ſhankes and yellow chappels ſculls: Or bid me go into a new made graue, And hide me with a dead man in his graue,

Juliet in *Romeo and Juliet*, 1623

Textual overlap in time: gap 1640–c1660



Interregnum 1649-1660: theatres closed
Restoration of monarchy 1660: theatres begin to flourish

Take-home messages & future research

Highlights:

- **automatic identification of textual overlap**: a new approach for finding potentially interesting documents in very large text collection
- provides a **starting point** both for traditional humanities research and for automatic analysis (e.g. network analysis)

Future:

- from “**what has been recycled**” to “**why**”
- towards much more robust methods that better tolerate variation
- the methods will work in future studies with similar kinds of data

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].

Hopper, Andrew, 2013. “Pamphlets and propaganda. Parliament versus the king in the 1640s”. History West Midlands. Available from: http://historywm.com/wp-content/uploads/issues/issue3/pdf/pp15-17_Hopper.pdf [Accessed on 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.