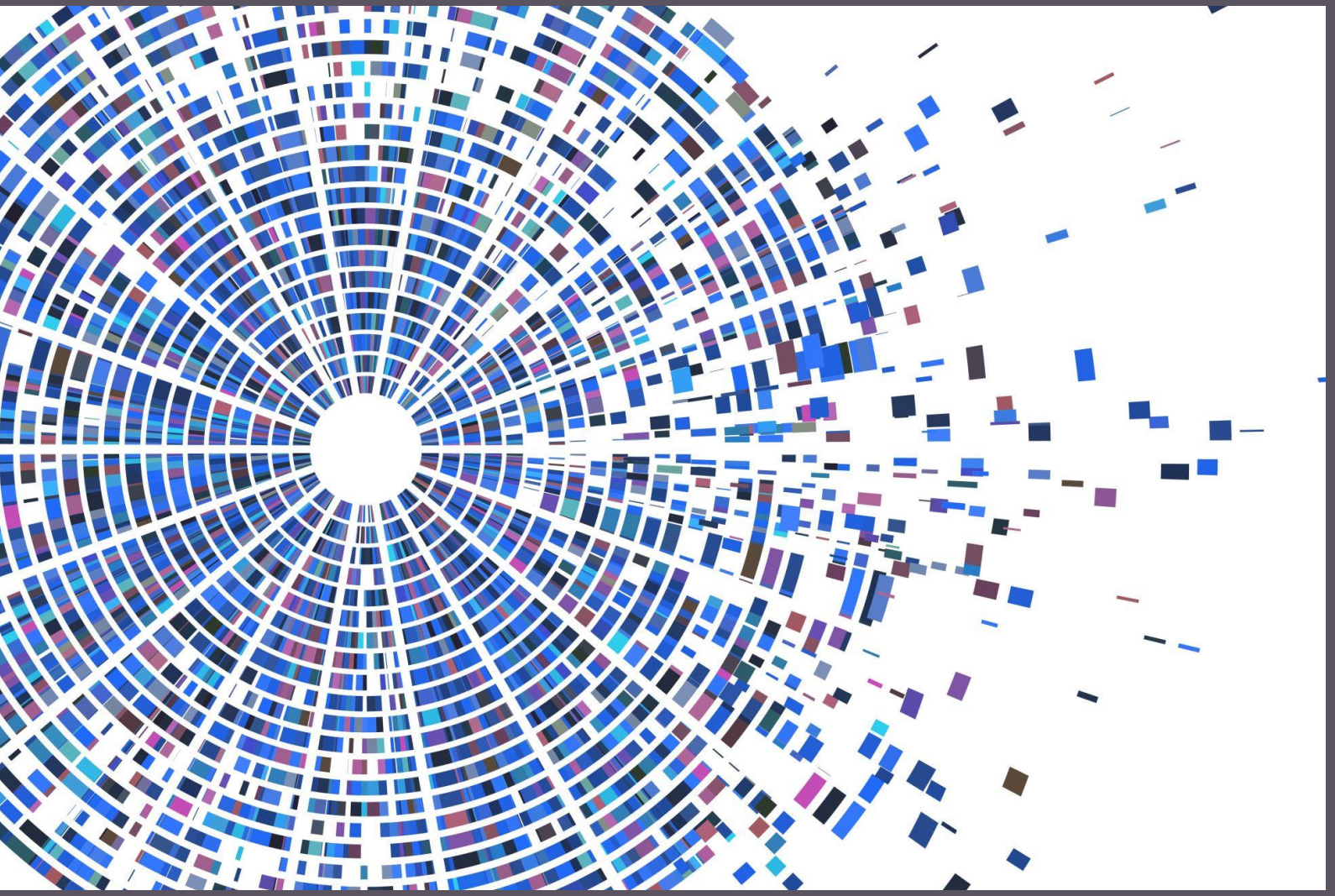


Lessons learnt from exploring *Cyfraith Hywel* through a digital lens

Dr Zoe Bartliff
Zoe.Bartliff@glasgow.ac.uk

Research Associate
Information Studies,
University of Glasgow

NLW Peniarth 28: Hawker - https://commons.wikimedia.org/wiki/File:Laws_of_Hywel_Dda_(f.4.r)_Hawker_cropped.jpg

Background

University of Glasgow

- *Cyfraith Hywel* – Medieval Welsh law text **first extant in the 12th century**, but some elements ostensibly dating to the 10th century
- **Foundation stone** of Welsh culture and identity in this turbulent period
- Approximately 80 extant copies, 40 dating from the period in which the law was active
- Each copy is at the **core the same text** but there are some distinct and **significant divergences** between the manuscripts (vocabulary, structure, content).
- Manuscripts usually grouped into one of three sub-traditions, namely **Iorweth**, **Blegwrydd** and **Cyfnerth**.

The Challenges

- Research into the text has reached a methodological impasse. There is a need to examine the material holistically, but current presentations of the material do not permit this.
- Text is either aggregated, losing detail, or sampled, limiting the broader applicability of the conclusions.
- Medieval Welsh, like many medieval languages, has few resources for automatically processing material.

The Aims

- Inspired by comparison between issues facing *Cyfraith Hywel* and those seen in relation to large scale data sets (cf. Busch 2014).
- Leverage the range of benefits possible with computer analysis to facilitate:
    1. efficient and effective access to the texts
    2. transparent and expandable research.

Research Methods

Pre-processing method –

- Analysis of key structural and semantic features.
- Manual application of XML encoding to 21 texts of *Cyfraith Hywel*
- Metadata overlay includes lemma form, grammatical breakdown, hypernym categorisation and structural features
- Makes the text computer readable

Analysis methods –

- Statistical language processing (drawn from corpus linguistics and NLP)
- Examples of the range of possible computational methods allowed by the encoding.

In practice…

Three incarnations of Aberystwyth, National Library of Wales, Peniarth MS 38:

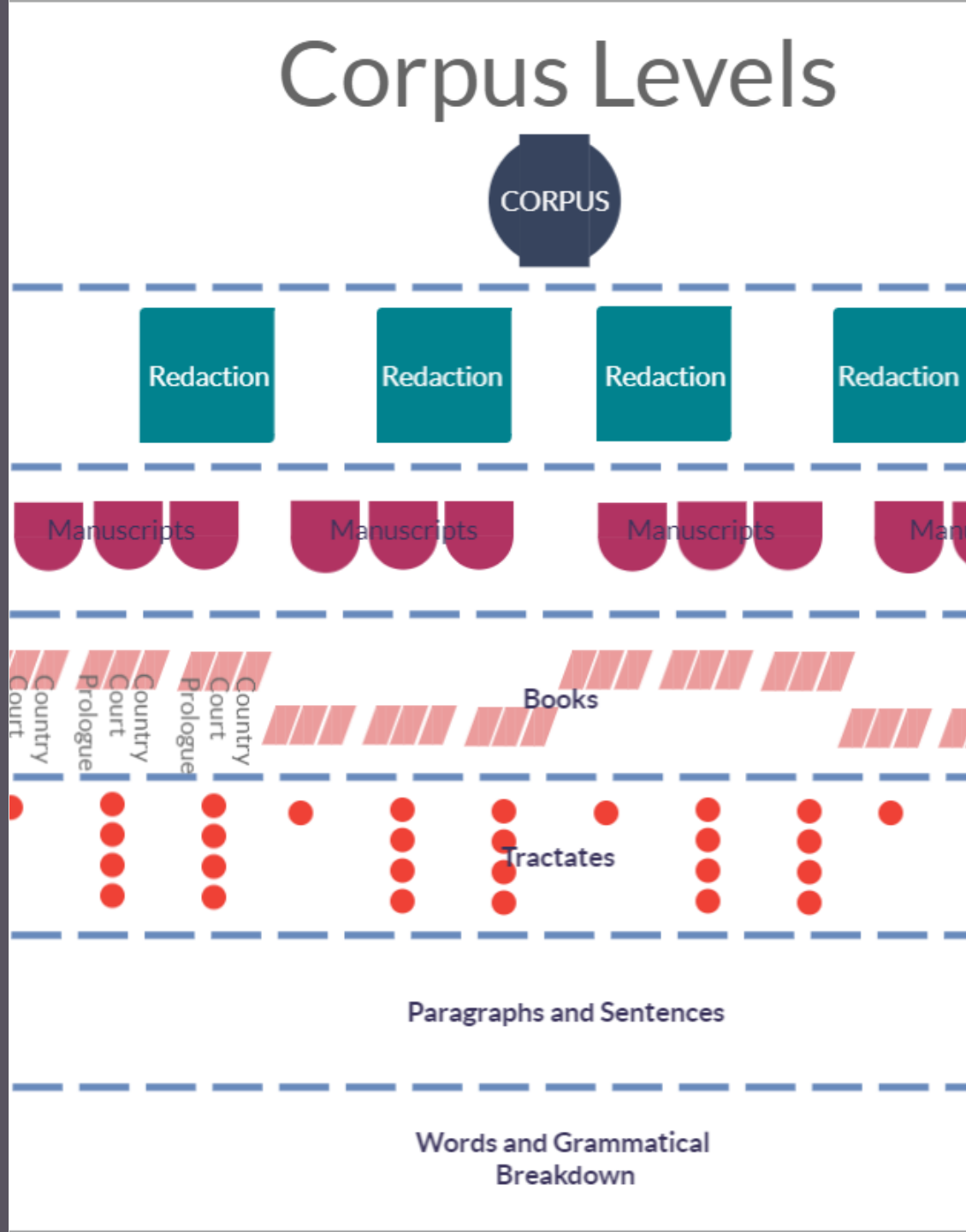
I. an image of the original manuscript digitised by Llyfrgell Genedlaethol Cymru (The National Library of Wales)

II. a screen grab from the *Rhyddiaith Gymraeg 1300-1425* project website

III. a screen grab of the computer and human readable XML encoding applied to the text through the process of creating the corpus. Uses an adapted from of TEI encoding.

Corpus Design

- Nested structure.
- Reflects different types of segmentation evident in the original manuscripts.
- Allows comparison at a variety of levels of granularity.

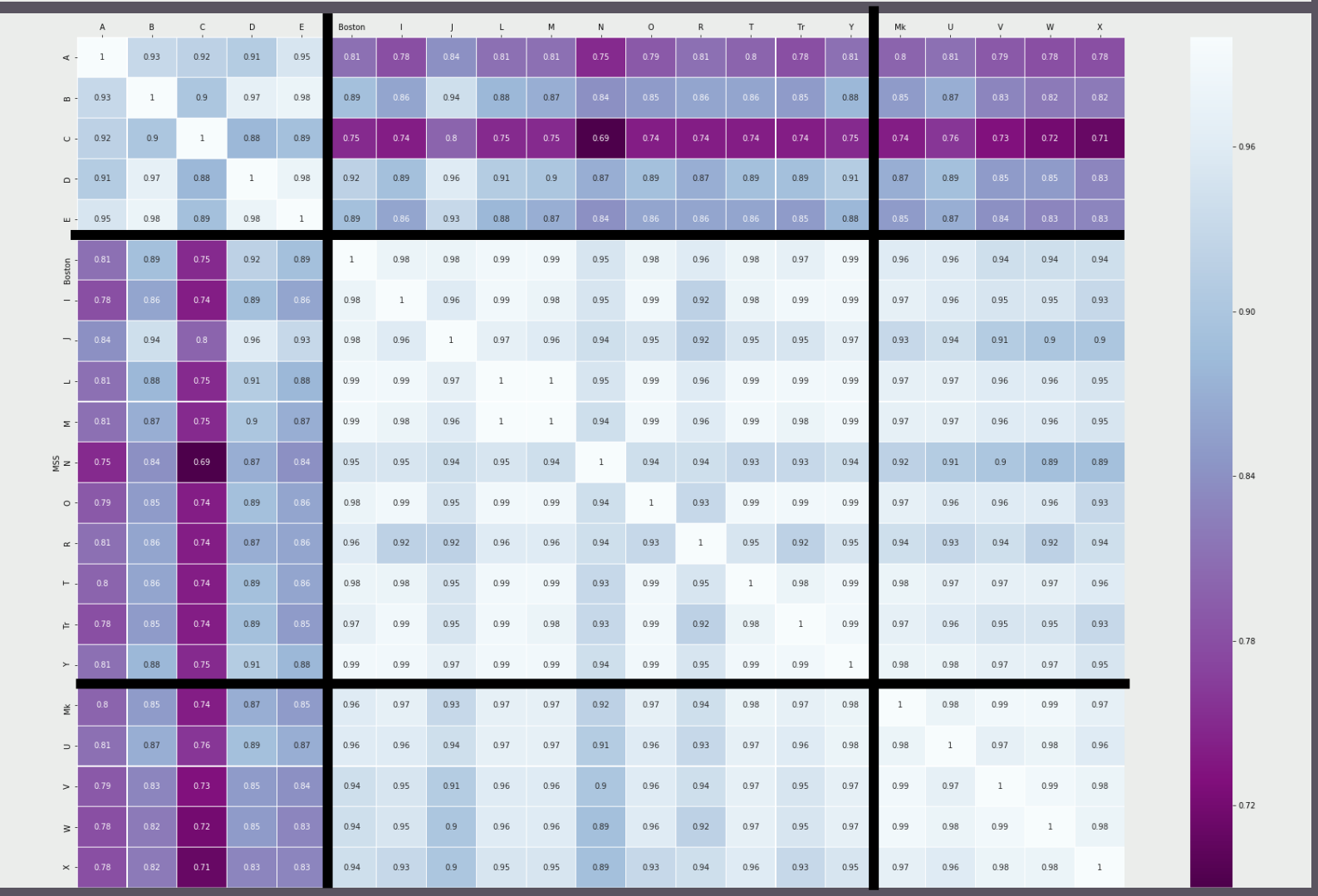


Research opportunities

Opens the text to statistical language processing.

Techniques explored include:

- token and lemma frequency analysis and comparisons
- unique token and lemma counts for language variety
- **measures of similarity and difference between manuscripts and sections (euclidean document term matrices and cosine similarity term matrices)**
- keyword analysis (relative frequencies and dispersion values)



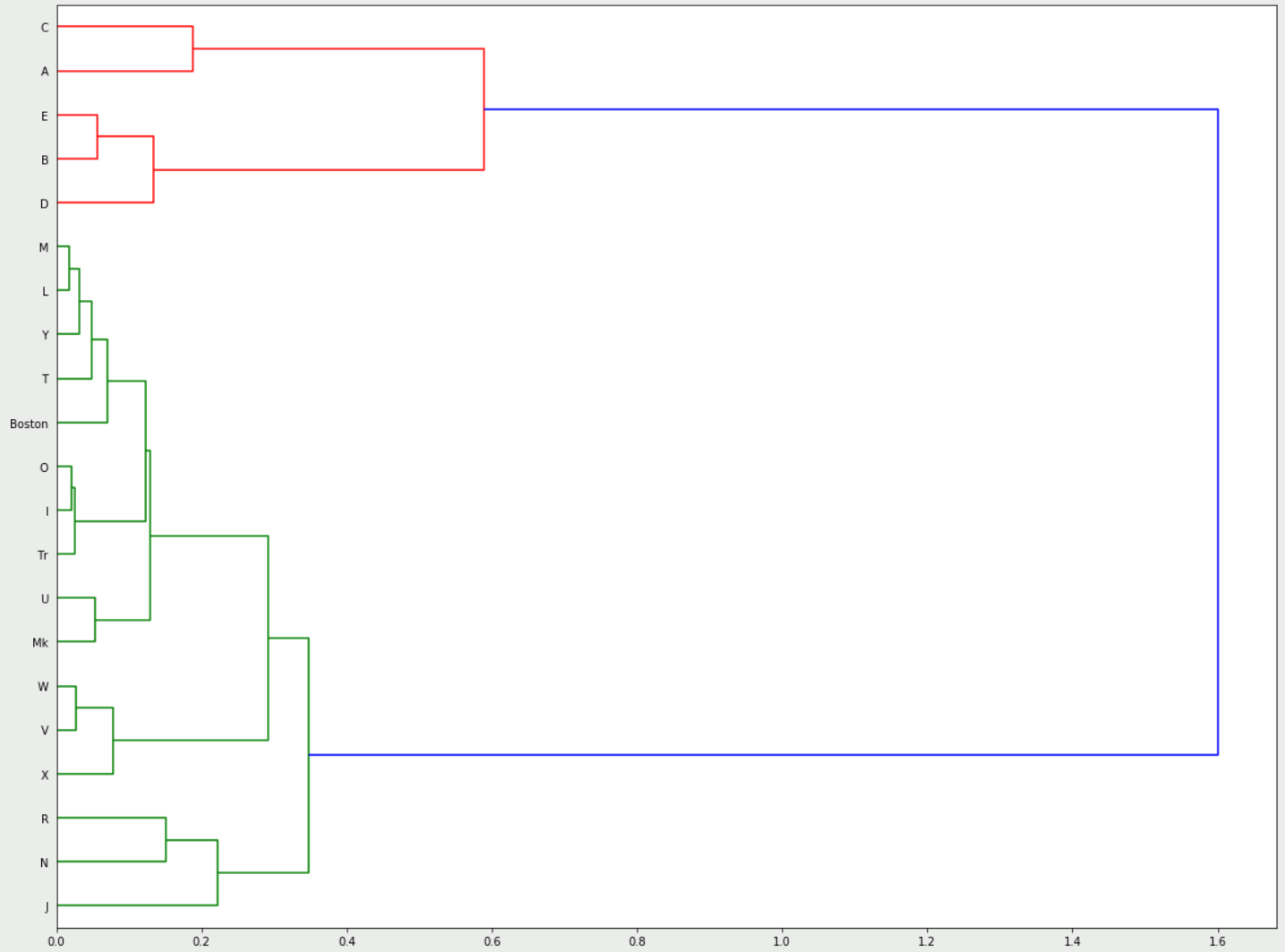
Cosine Similarity matrix with heatmap to indicate extent of similarity between manuscripts.
Sections indicate MSS belonging to the Iorweth, Blegywryd and Cyfnerth redactions respectively.

Cosine Similarity Matrix

- Empirical measure of similarity
- Calculated by measuring the Euclidian distance of each word from each other word and then inversed for the measure similarity.
- Applied here to the whole text, but can be applied on the book, tractate or even sentence level.
- 1 is complete similarity, 0 complete difference

Hierarchical Clustering

- Based on measures of similarity
- Similar to manuscript stemma in form.
- Distinct groupings between the Iorweth tradition and the other two, but Cyfnerth and Blegwryd more interwoven.
- Clear clusters of similarity
- Highlights points of divergeance.



Hierarchical clustering dendrogram to indicate branches of similarity between manuscripts.

What can this tell us?

- The division of the tradition into redactions is, on the whole, supported but the relationships between the manuscripts within each redaction is more complicated.
- Most measures displayed here suggest that the Cyfnerth redaction is the most homogeneous and the Iorweth the most diverse.
- The patterns of similarity vary between the `Book' level divisions of the manuscripts, suggesting that different sections of law may draw inspiration from divergent branches of the overarching tradition. This aligns with recent trends in scholarship to consider manuscript production as `cross-fertilization' (Jenkins 2000, p.13) rather than a direct line of descent.

Key Thought

‘Thus, the ideal role of the computer and the purpose of computing in digital humanities is not to make research better, faster and/or cheaper. On the contrary, as a number of writers have argued, computing should be about making problems more difficult, more complex, more thrilling — computing is, or can be, ‘a telescope for the mind' (Masterman, 1962).’ (Ortolja-Baird 2019)

Conclusions and future work

The encoding is a foundational step towards accurately incorporating digital humanities methods with medieval texts.

Could incorporate (semi)automatic analysis using manually encoded texts to train machine learning systems

Statistical analysis gives a shared language from which to make judgements.

It reveals patterns that would be unseen through manual analysis

Best used in conjunction with manual methods, as the 'human' element can be difficult to determine

Plans to:
- explore corpus more fully at book & tractate level
- make corpus open access
- Explore viability for cross-language comparison

Thank You!

If you have any thoughts or questions,
I'd love to hear from you:

Zoe.Bartliff@glasgow.ac.uk