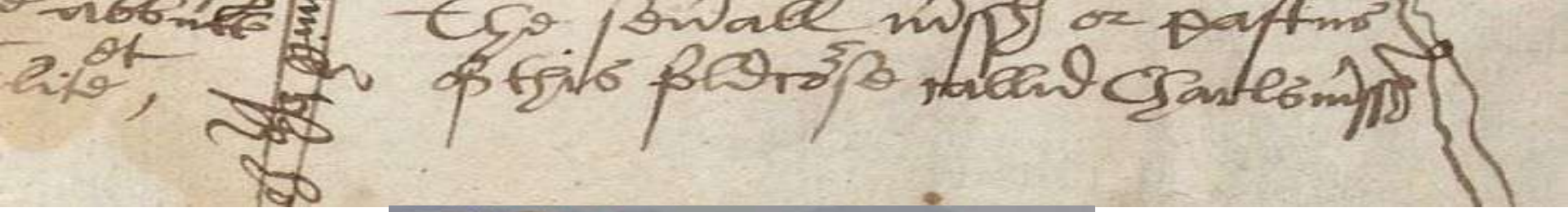
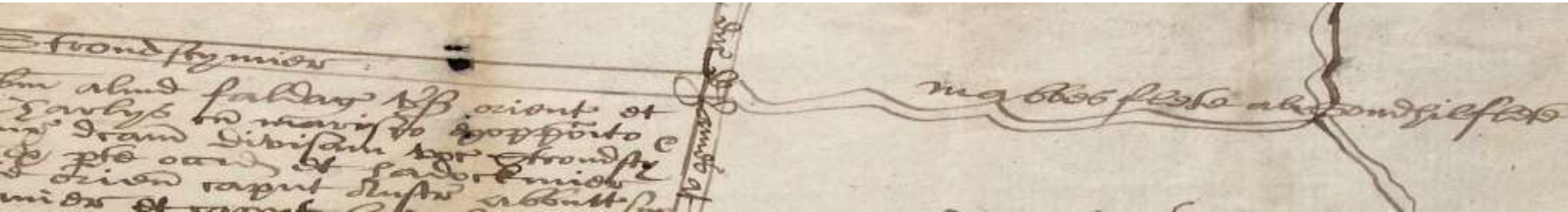


DIGGING INTO DATA CHALLENGE

ChartEx: Discovering spatial descriptions and relationships in medieval charters

Sarah Rees Jones, Christopher Power
Data Management and Data Mining Research Workshop
University of York
15 Nov 2012

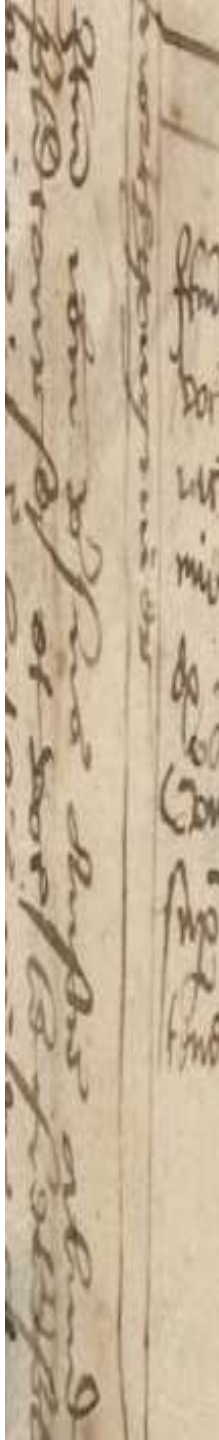


1. Project overview

- DIGGING INTO DATA CHALLENGE

- “As the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these everyday materials. Digging into Data challenges the research community to help create the new research infrastructure for 21st century scholarship. ”

- UK : AHRC, ESRC, JISC
- US: NEH, IMLS, NSF
- CAN: SSHRC
- Netherlands Organisation for Scientific Research

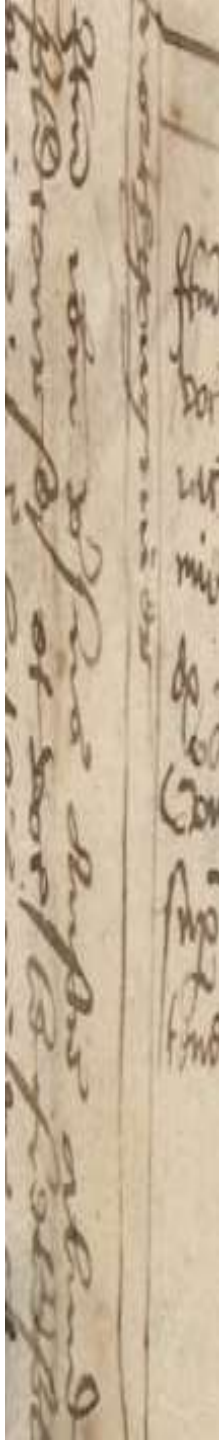


ChartEx Project Team

- ***University of York.*** History, Human Computer Interaction
- ***University of Brighton.*** Natural Language Processing
- ***University of Leiden.*** Data Mining
- ***University of Washington.*** History, Web Services
- ***University of Toronto.*** History and Digital Archives
- ***Columbia University.*** History and Digital Libraries

- ***Data Repositories:*** The National Archives (UK), Borthwick, DEEDS Project Uo Toronto, Columbia Digital Humanities

- ***Advisory Panel:*** various cognate projects

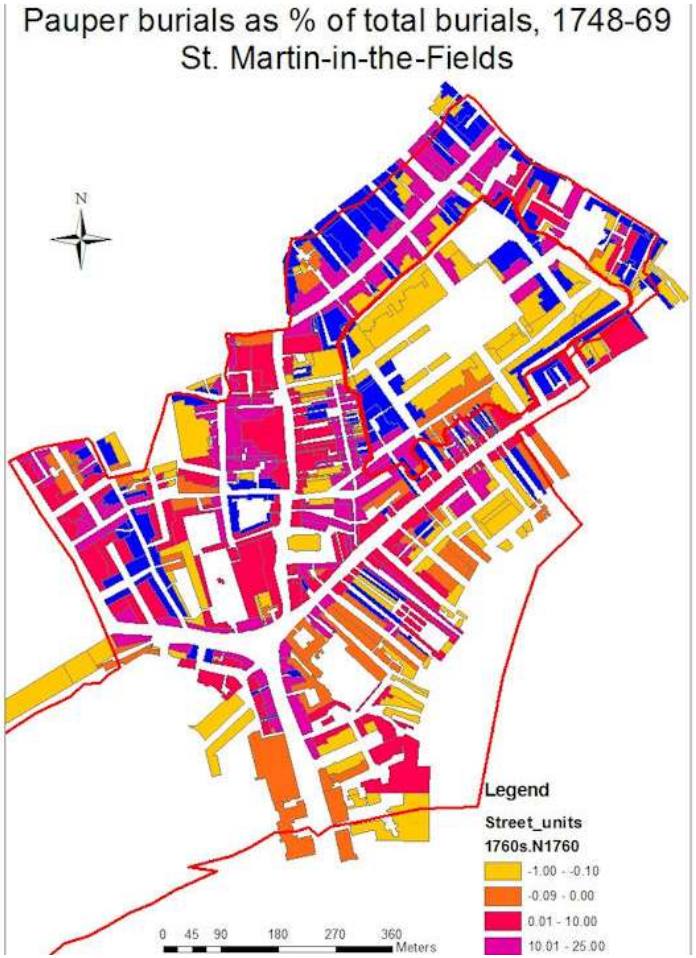


Why Charters?
Stories about people and places

- **408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. at Christmas. Warranty. Seal.**
- **Witnesses: Geoffrey Gunwar, William de Gerford[b]y,' chaplains,Robert de Farnham, Robert le Spicer, John le plastrer, Walter de Alna goldsmith, Nicholas Page, Thomas talliator, Hugh le bedel, John de Glouc', clerks, and others.**
- **January 1252 [1252/3].**
- **SOURCE: VC 3/Vi 326 (161 mm. x 137 mm.)**
- **ENDORSEMENT: Petergat', Donacio facta vicariis de domo que fuit Thome aurifabri; Simonis Evesham.**
- **SEAL: Slit.' Hole in MS.**
- **NOTE: See 403.**

Why Charters?

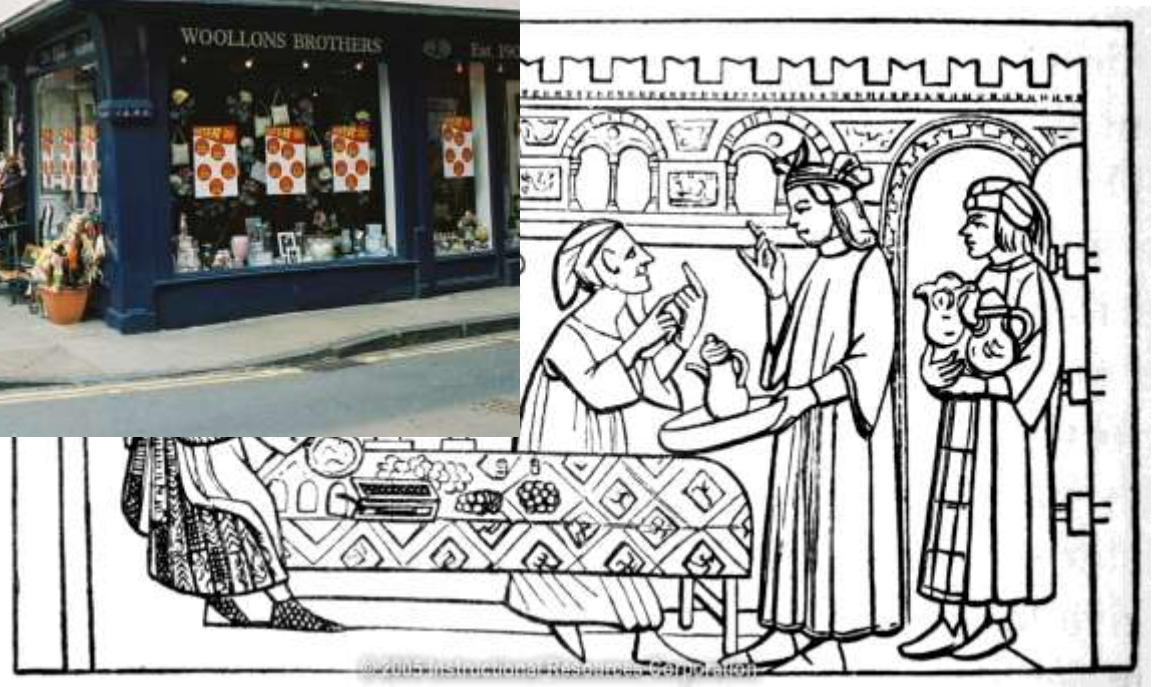
History: the analysis of social, cultural and economic networks



- First establish the locations in which people lived
- Then map other data (here about poverty) on to those patterns.
- ChartEx – going beyond the manual capture of data.

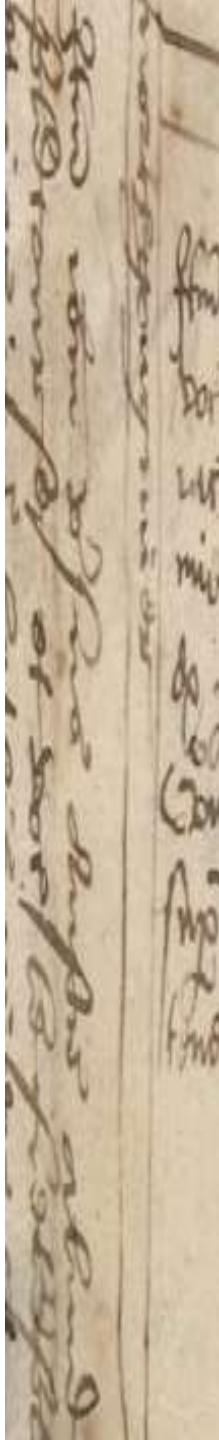
Why Charters?

Heritage: going beyond buildings to people and lived environments.

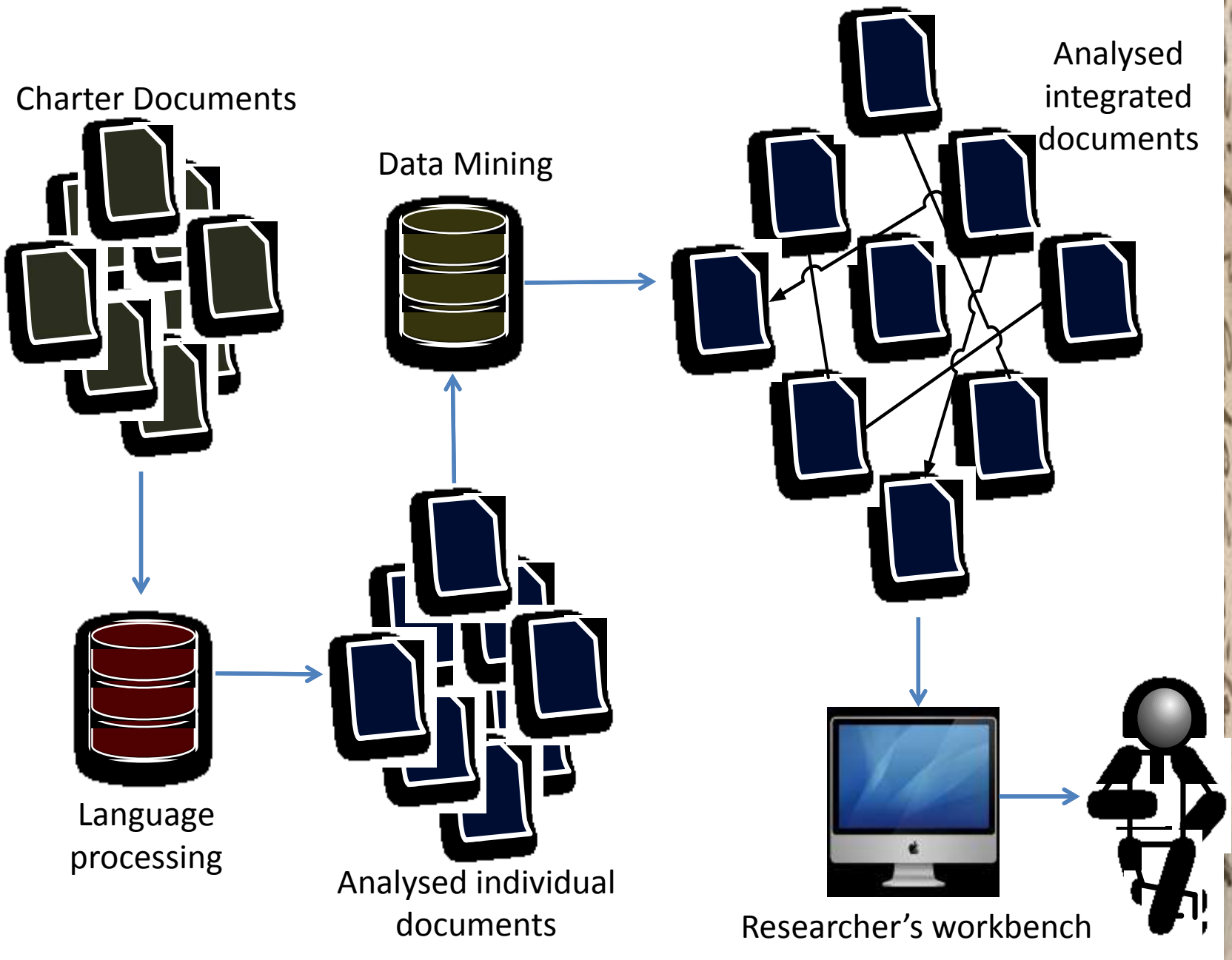


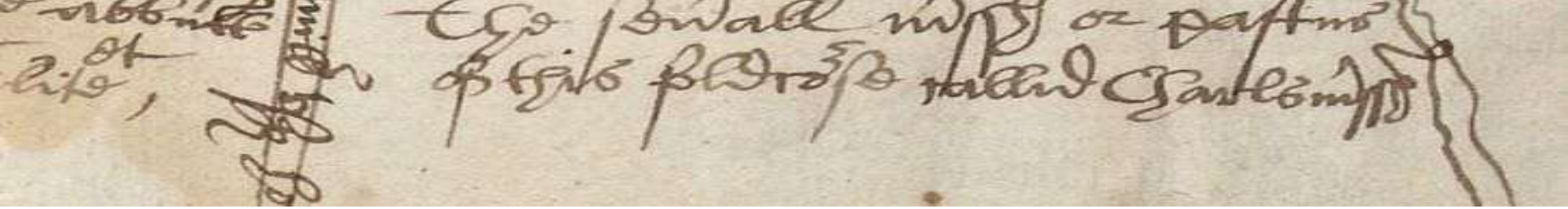
ChartEx: Digging with Technology

- Automatically identify key parts of documents and extract them using natural language
- Automatically identify new relationships and knowledge through data mining
- Allow users to explore and understand the information available with user interfaces that support the types of complex tasks professional researchers do



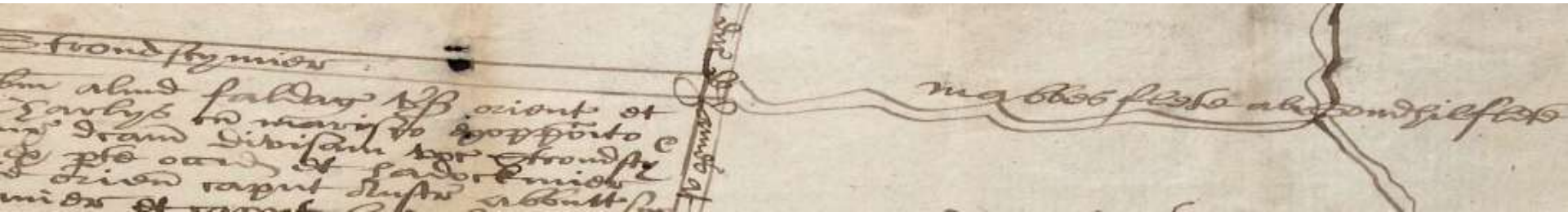
ChartEx architecture





2. Manual markup

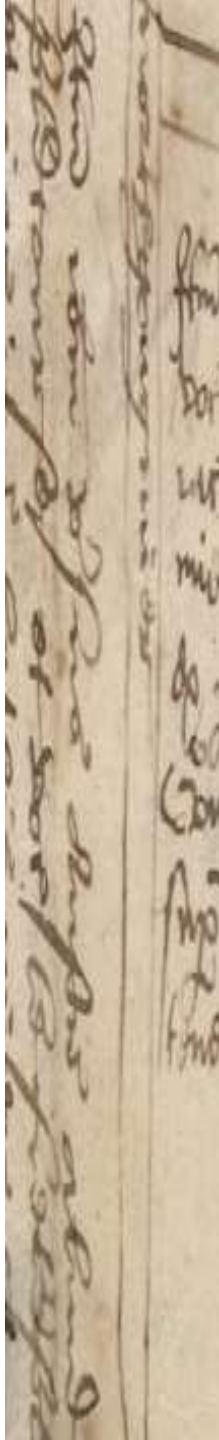
Training data for NLP and DM



Step 1a: Identify Entities

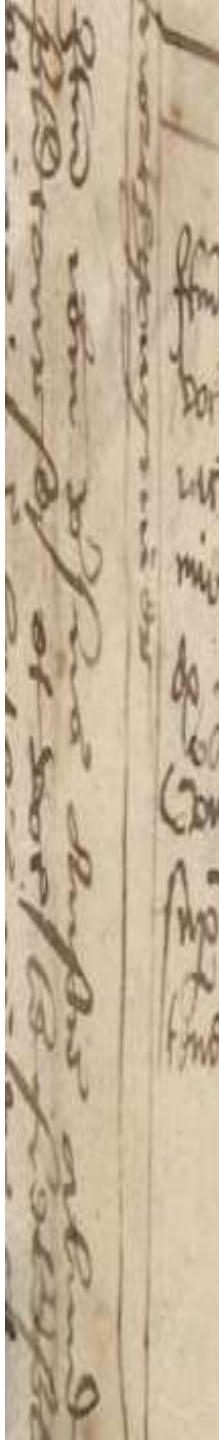
- Initial examination charters for entities about actors, sites and events…….

- 408. Grant by **Thomas** son of **Josce goldsmith** and citizen of York to his younger son **Jeremy** of half his **land** lying in length from **Petergate** at the **churchyard of St. Peter** to **houses** of the **prebend of Ampleford** and in breadth from **Steyngate** to **land** which **mag. Simon de Evesham** inhabited; paying **Thomas** and his heirs 1d. or [a pair of] white gloves worth 1 d. at **Christmas**. Warranty. Seal.

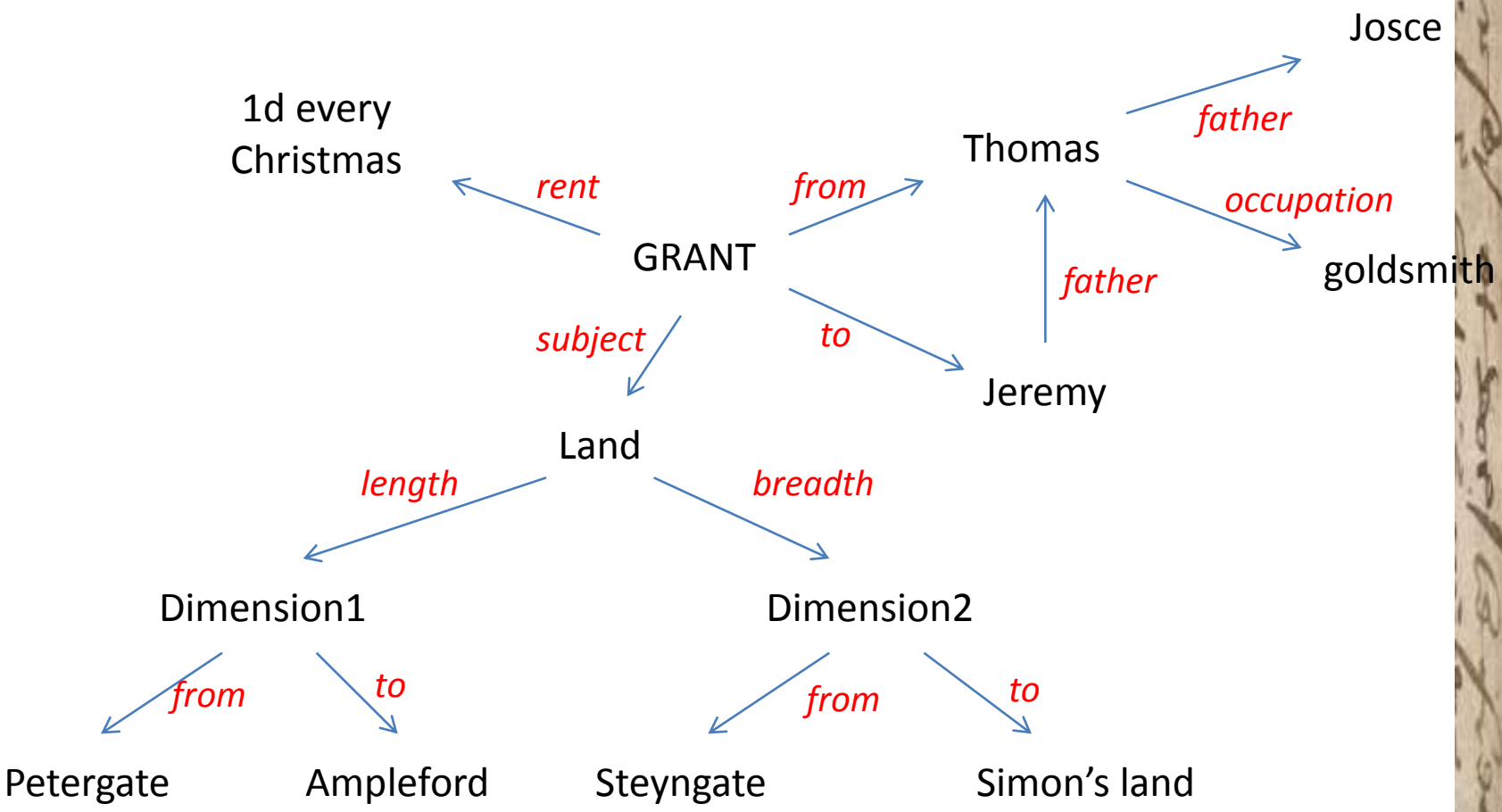


Step 1b: Identify Relationships between Entities

- Initial examination of charters for relationships between entities …
- 408. **Grant by** Thomas **son of** Josce goldsmith and citizen of York **to** his **younger son** Jeremy of half his land lying **in length from** Petergate at the churchyard of St. Peter **to** houses of the prebend of Ampleford and **in breadth from** Steyngate **to** land which mag. Simon de Evesham **inhabited**; **paying** Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. **at** Christmas. Warranty. Seal.



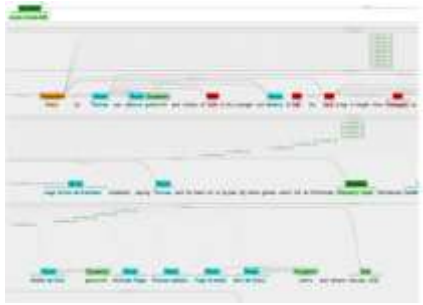
Representing relationships



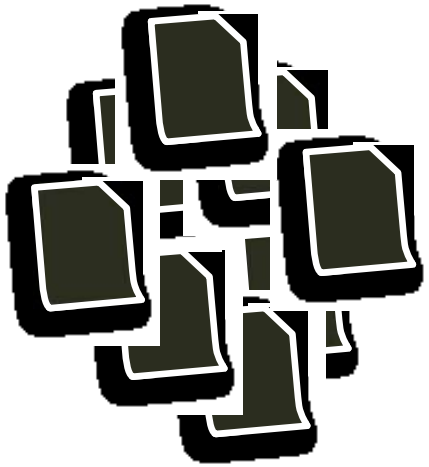
Step 1c: BRAT tool

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. (http://brat.nlplab.org/)

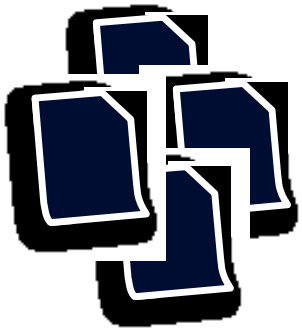
NLP Training Methodology



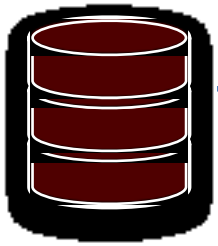
Unmarked Documents



50 documents

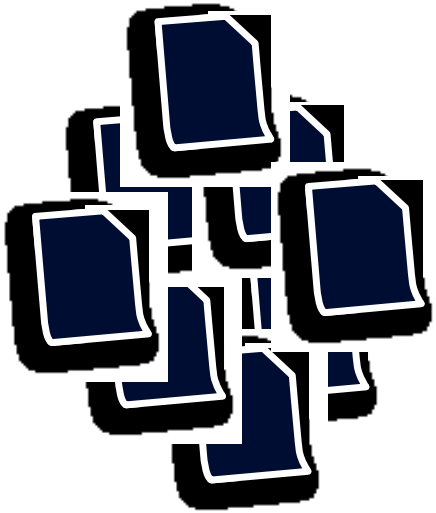


Training documents

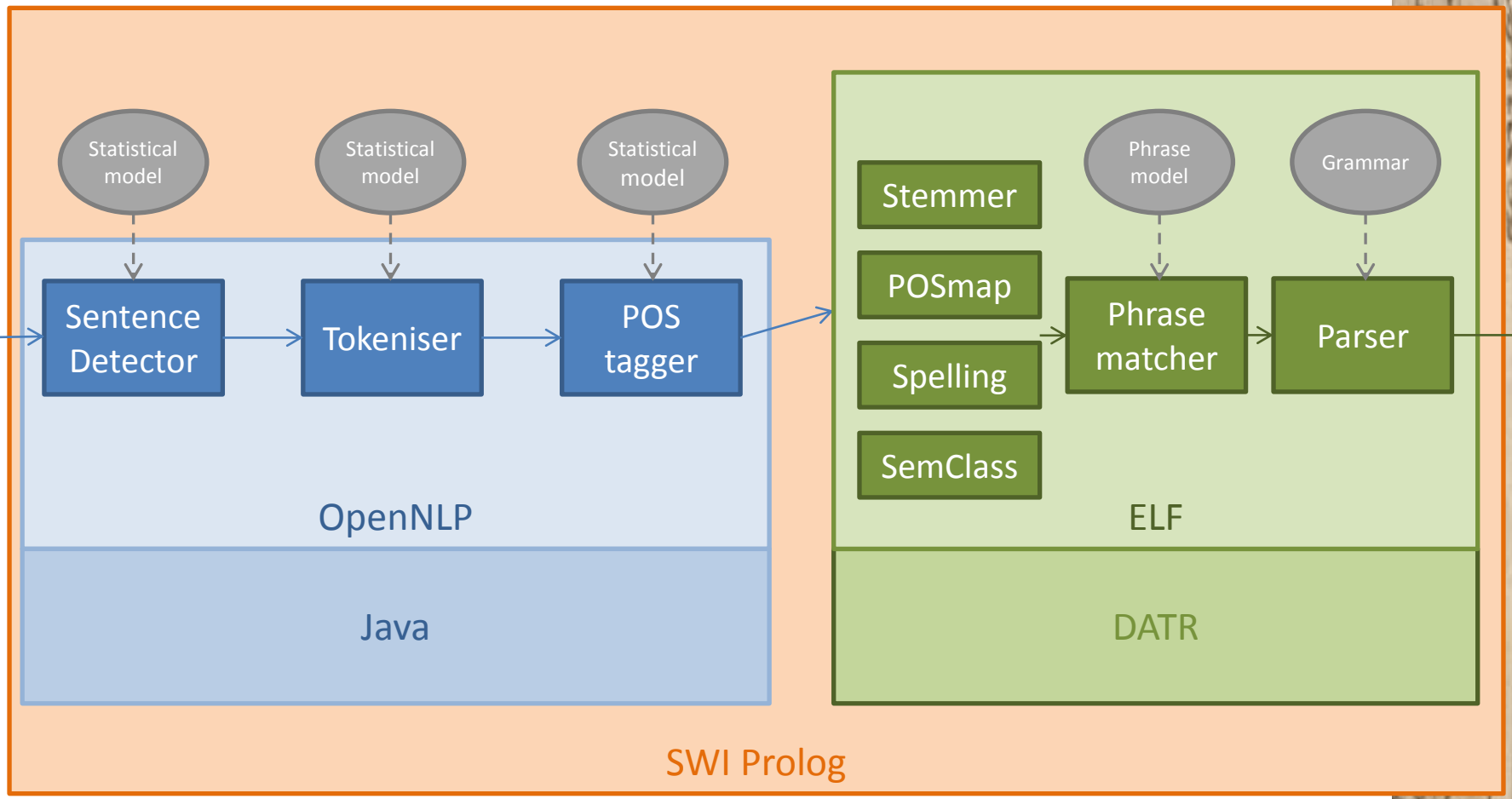


NLP

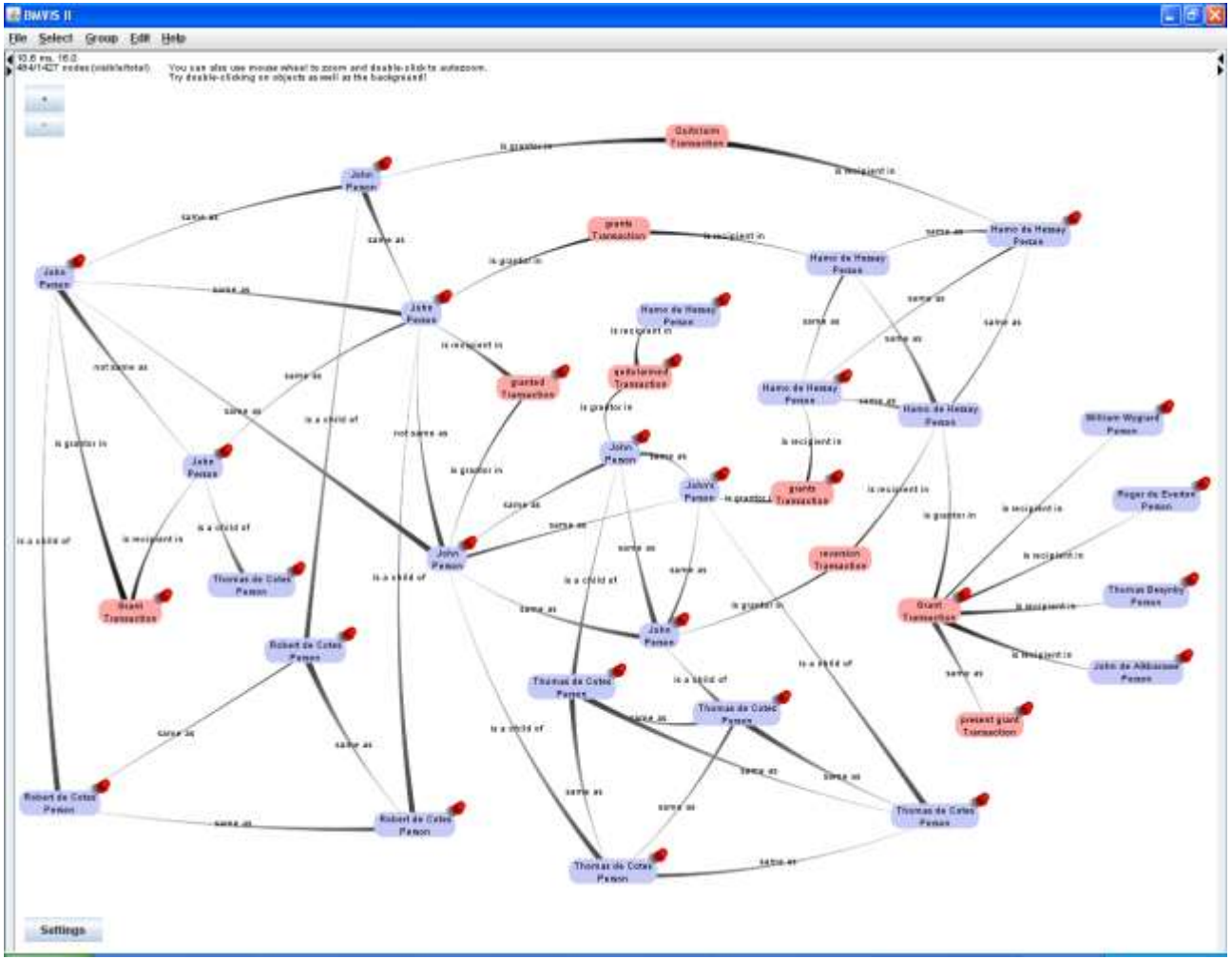
Newly marked documents

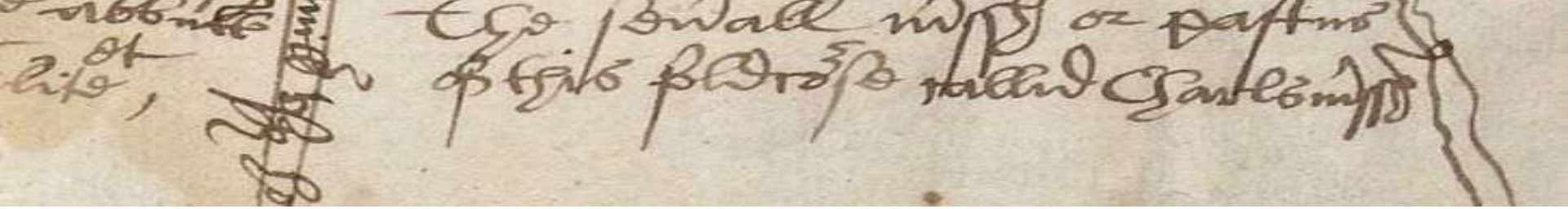


ChartEx NLP architecture (work in progress)

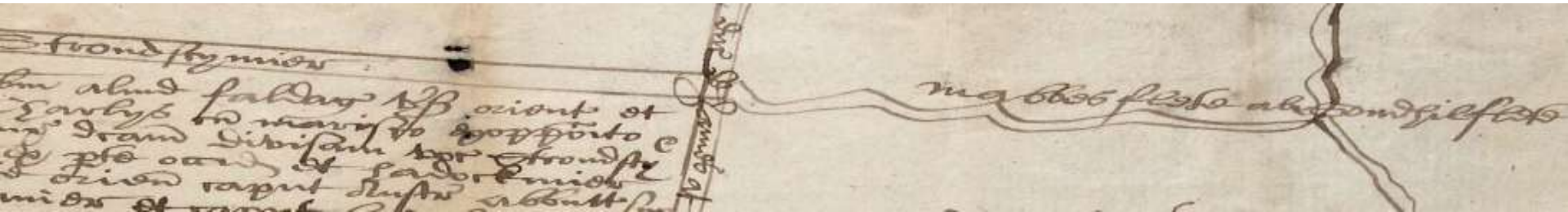


Data Mining
(work in progress)



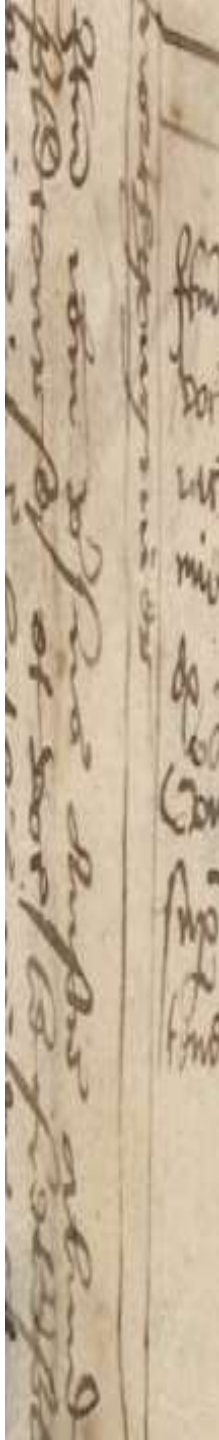


3. Interaction



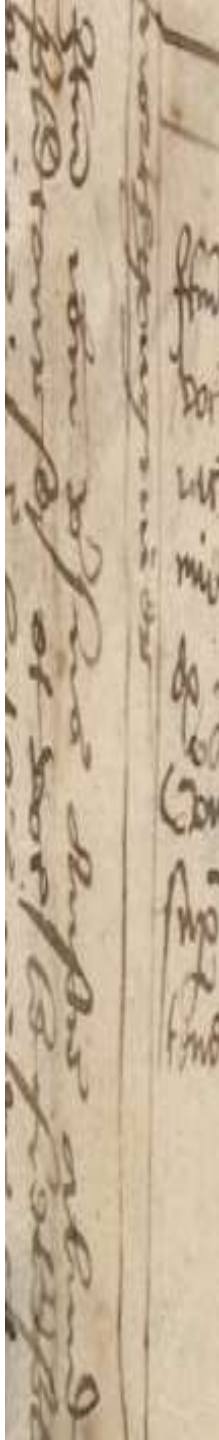
Virtual Workbench

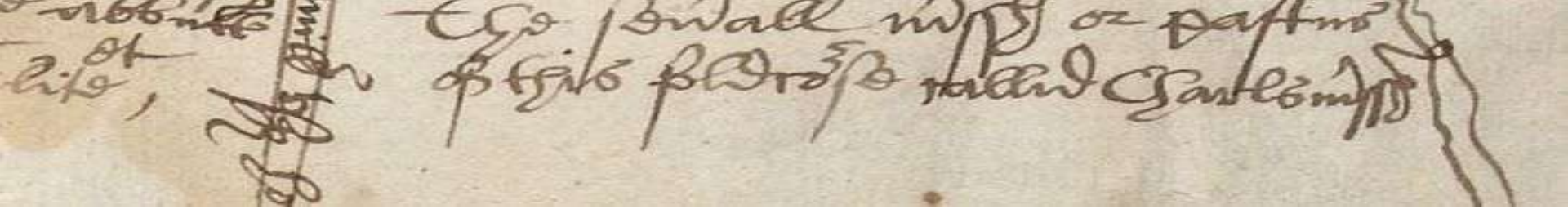
- Contextual inquiry conducted with 8 historians
  - "Bring your own charters" sessions
  - Studied how historians work with documents, the ways they construct their mental models of a problem and the types of information they record



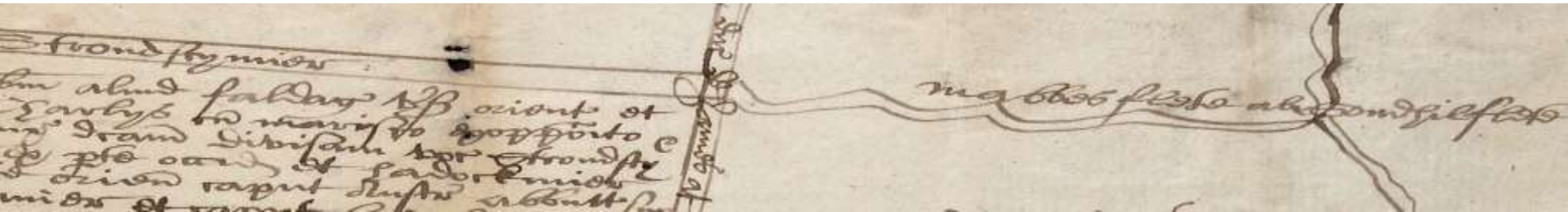
Virtual Workbench (2)

- Transcription and analysis identified three distinct stages of work:
  - Working with collections – finding documents of interest
  - Investigating relationships – exploring a small set of documents for particular entities
  - Building models – creating own view of data contained in documents, sometimes adding external knowledge
- Currently most systems only support the first step well, and occasionally the second step





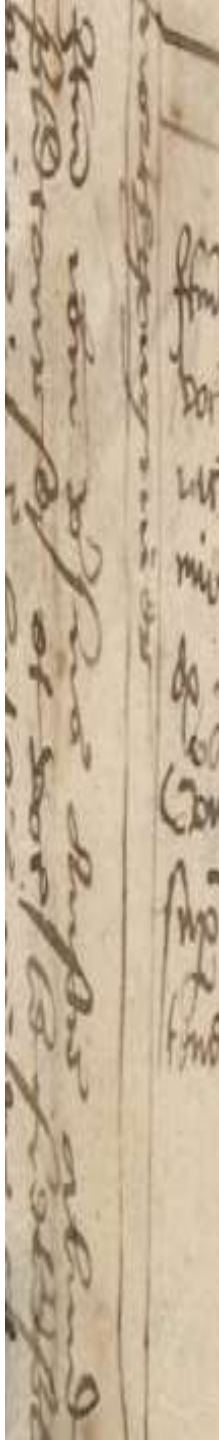
4. Challenges



Markup challenges

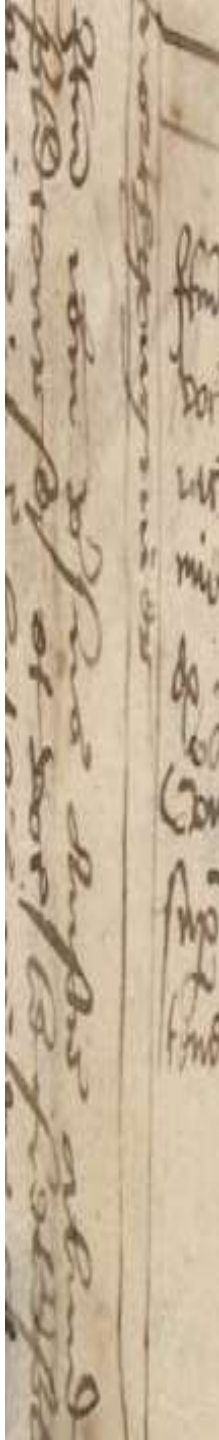
- Designing a markup scheme – we used the KANGA methodology (originally from the OS)
- Annotating documents – we used BRAT
- Interdisciplinarity – we used our HCI team to elicit the markup language from our Historians and applied it using a tool designed for Bio-NLP markup!
- Maintaining focus on task in hand was sometimes difficult
- Establishing good markup practices (inter-annotator agreement tests etc.) required effort.

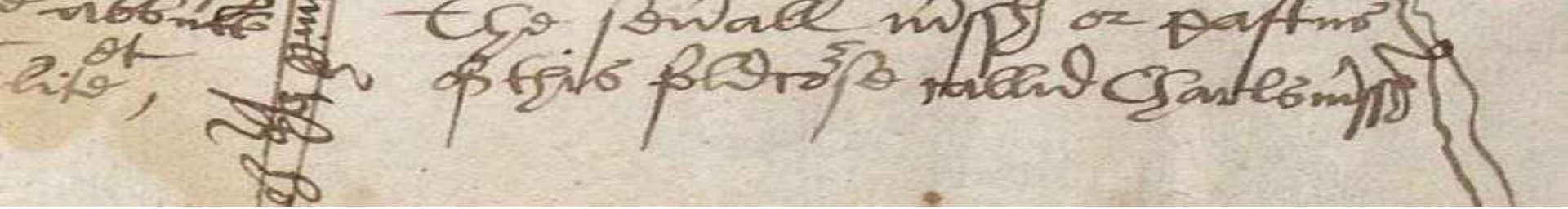
KANGA: http://sourceforge.net/apps/mediawiki/confluence/index.php?title=ROO-1.0:Kanga_Ontology_Construction_Methodology



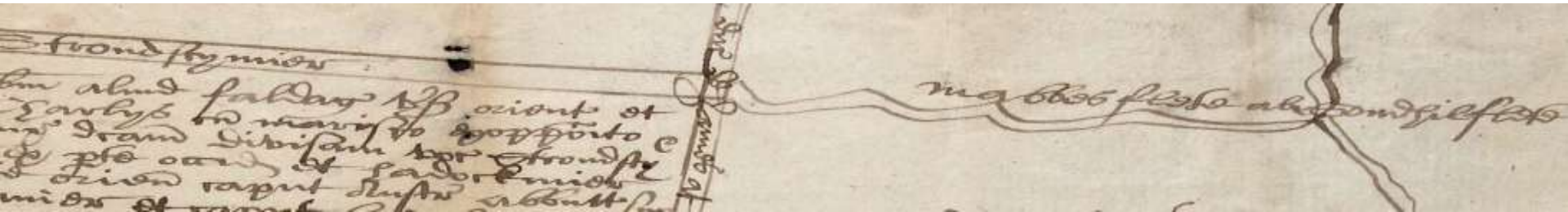
Issues with Charter data

- Medieval shorthand (especially of Latin)
- Status of ‘apparatus’ – editorial additions or corrections made over the years
- OCR errors (of course)
- Different markup conventions, or different interpretation of markup conventions such as EAD (Eg main transaction type – data or metadata?)
- Interesting relationships and mixing between Latin and English
- Information omitted during transcription – notably, locations. (TNA are putting them back for us!)





5. Conclusions



Conclusions

- Steady progress overall – markup phase completed successfully, technical work now coming into its own
- Interesting interdisciplinary challenges – methodologies, analytic approaches, vocabularies etc.
- Lots of positive engagement from all parties to keep us all sane!