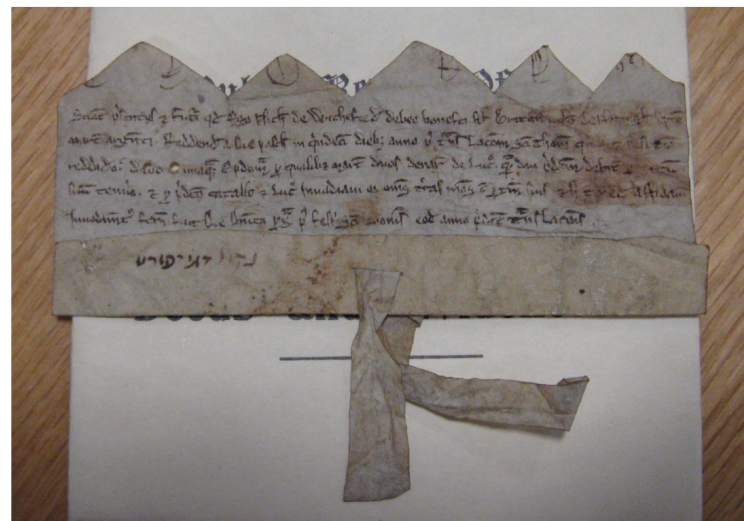


Exploring medieval charters: the ChartEx Project

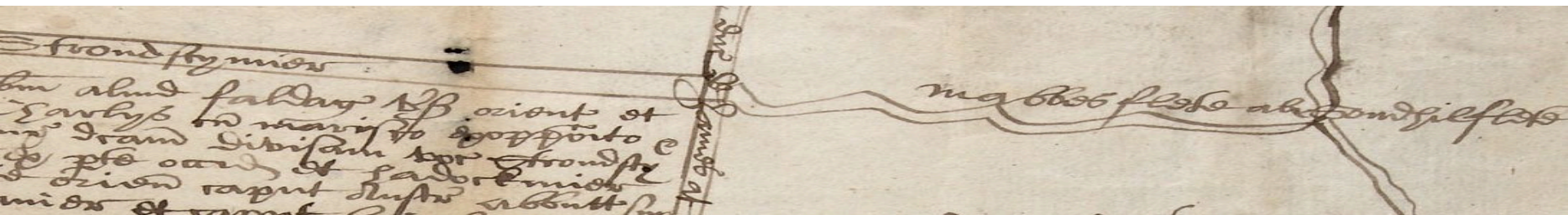
Roger Evans

Natural Language Technology Group
University of Brighton



www.chartex.org

@ChartExProject



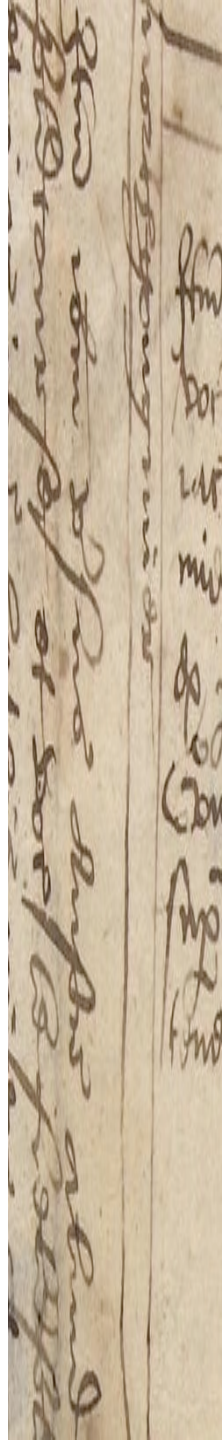


Charters

Charters are legal documents recording agreements about land and buildings

(like modern 'Title Deeds')

- Charters are among the oldest written records of 'everyday' European life and society – going right back to 1100's
- Early charters were all in Latin – also in English from around 1300's
- Thousands of charters survive, handwritten documents on delicate scrolls
- Charters contain lots of information about people, sites, occupations and social and economic relationships, but unlocking them is a highly specialised, manual task.
- Historians have invested a lot of effort recently transcribing charter texts to digital form.

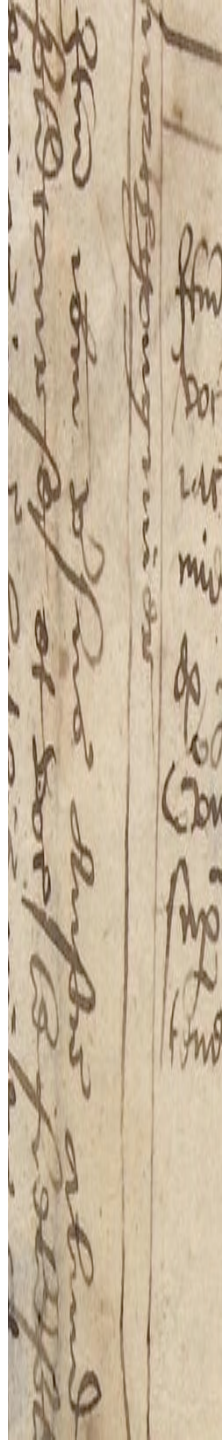


Example Charter

(from the Vicars Choral collection, University of York)

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.

NB: this is actually a translation from a Latin original.

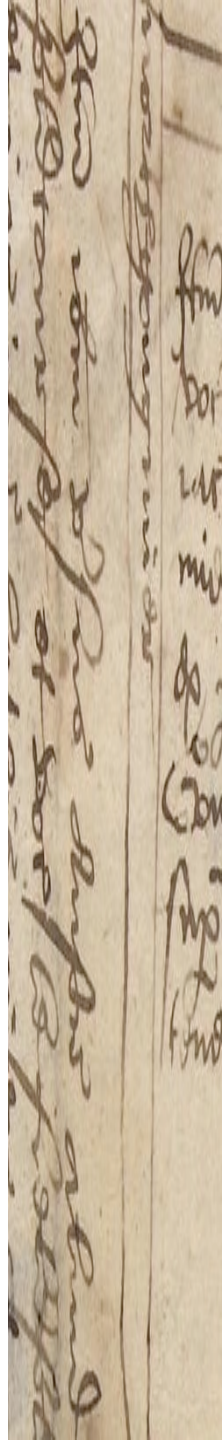


Example Charter

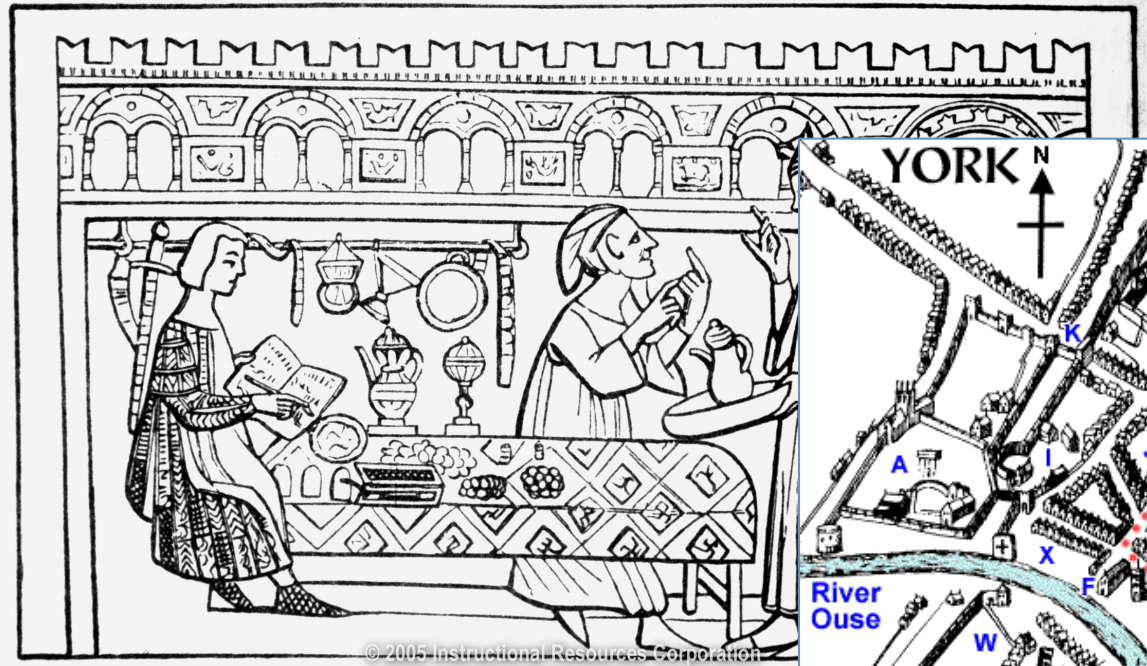
(from the Vicars Choral collection, University of York)

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.

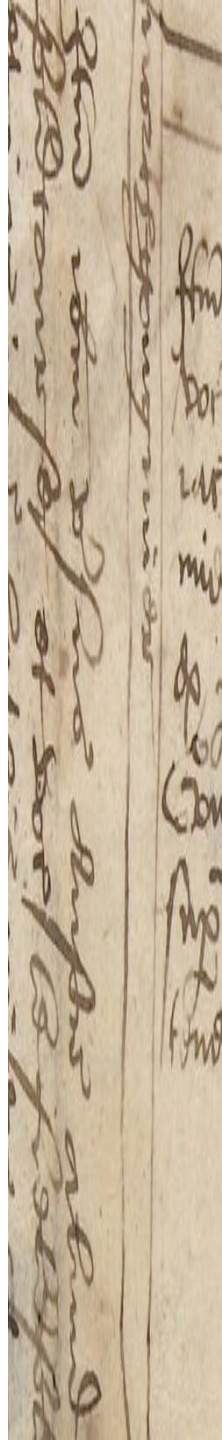
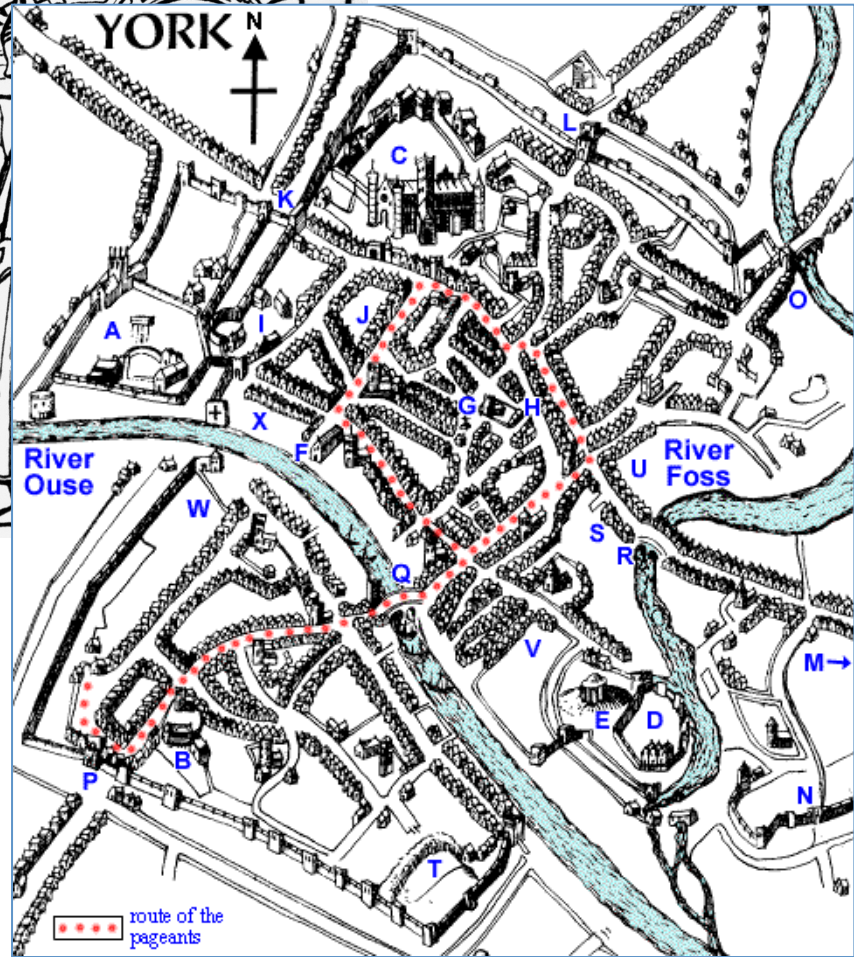
NB: this is actually a translation from a Latin original.



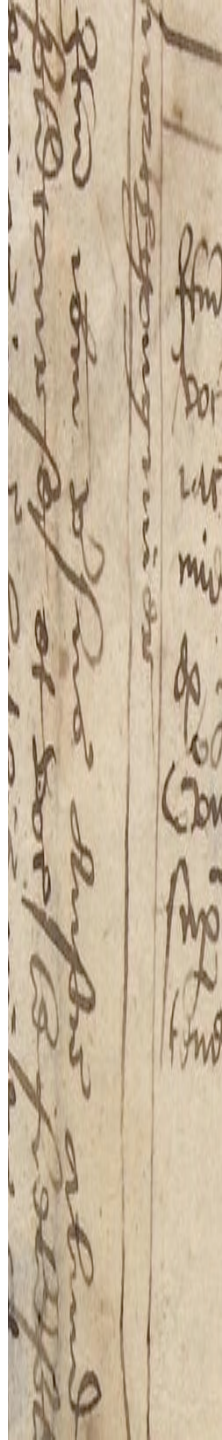
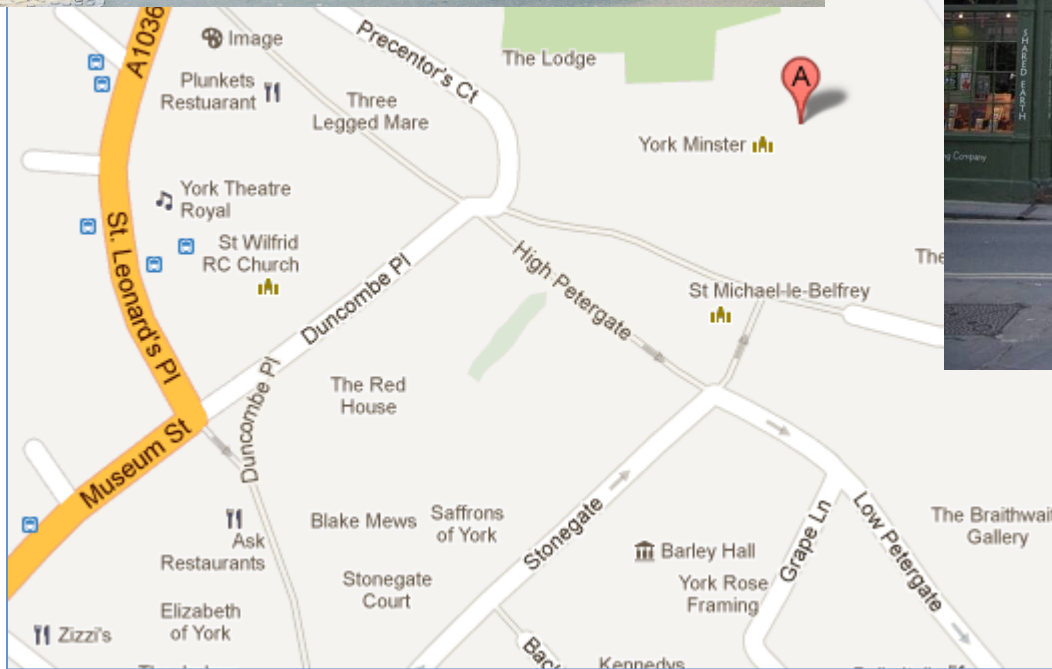
Urban Historic Topography



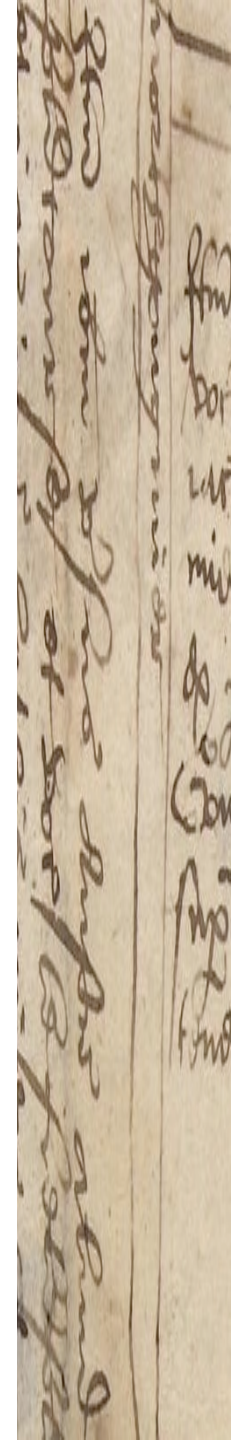
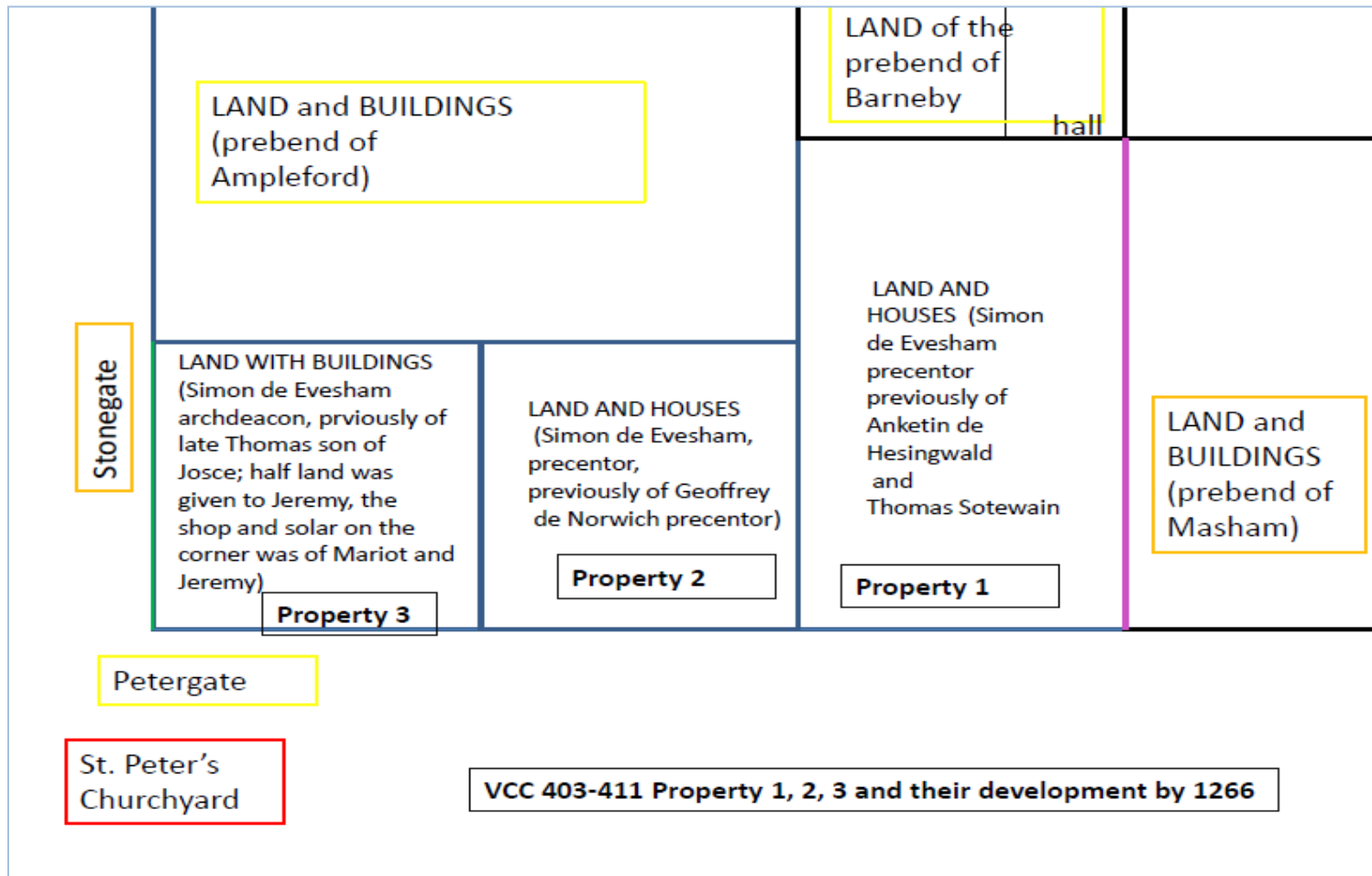
Medieval
Petergate
(York)



Petergate today



Conceptual plan of logical relationships



ChartEx project aims

- Apply Natural Language Processing (NLP) and Data Mining (DM) to medieval charters to extract information about places, people and transactions
- Develop a Virtual Workbench (VWB) for historical researchers to explore relationships across whole collections of charters



ChartEx project team

Partners

University of York: History (Sarah Rees Jones, Stefania Perring) and Human Computer Interaction (Helen Petrie, Christopher Power, David Swallow)

University of Brighton: Natural Language Processing (Roger Evans, Lynne Cahill)

University of Leiden: Data Mining (Arno Knobbe, Marvin Meeng)

University of Washington: History, Web Services (Robert Stacey, Jon Crump)

University of Toronto: History and Digital Archives (Michael Gervers, Robin Sutherland-Harris)

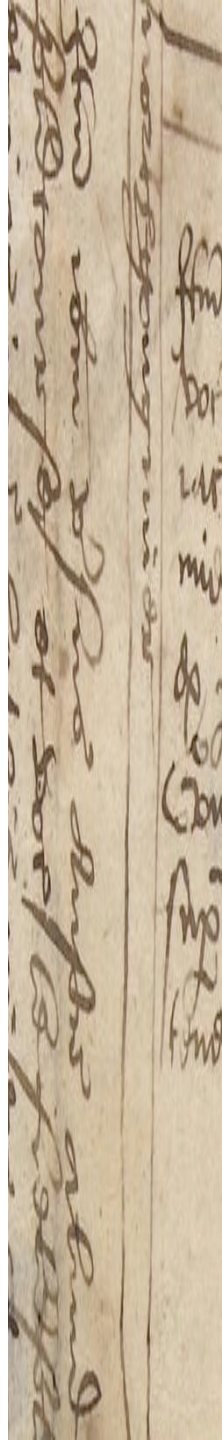
Columbia University: History and Digital Libraries (Adam Kosto)

Data Repositories: The National Archives (UK); Borthwick; DEEDS Project, U of Toronto; Columbia Digital Humanities (CBMA)

Funding

Digging into Data Challenge (www.diggingintodata.org)

AHRC, ESRC, JISC (UK); NEH, IMLS, NSF (USA); SSHRC (CAN); OSR (NL)



ChartEx development process



Kanga Methodology

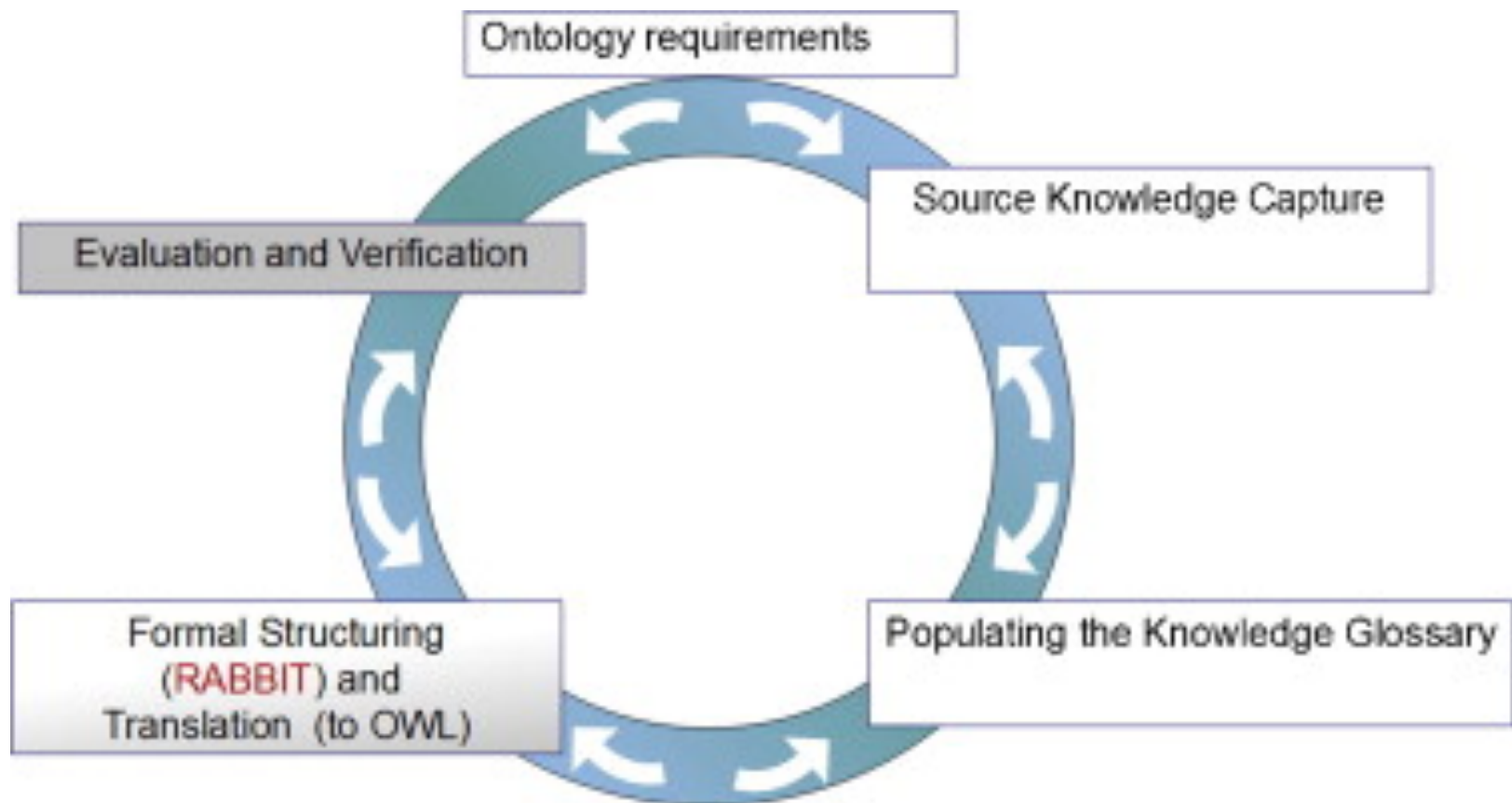


Fig. 1 The phases of the Kanga Methodology. The white boxes indicate the phases performed by domain experts. The formal structuring is done by domain experts using Rabbit, while the translation to OWL is ...

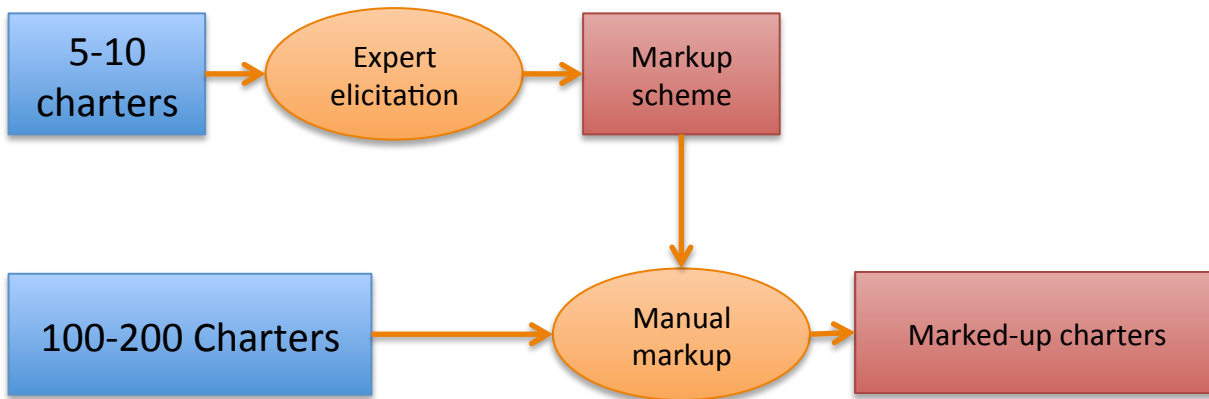
Ronald Denaux , Catherine Dolbear , Glen Hart , Vania Dimitrova , Anthony G. Cohn

Supporting domain experts to construct conceptual ontologies: A holistic approach

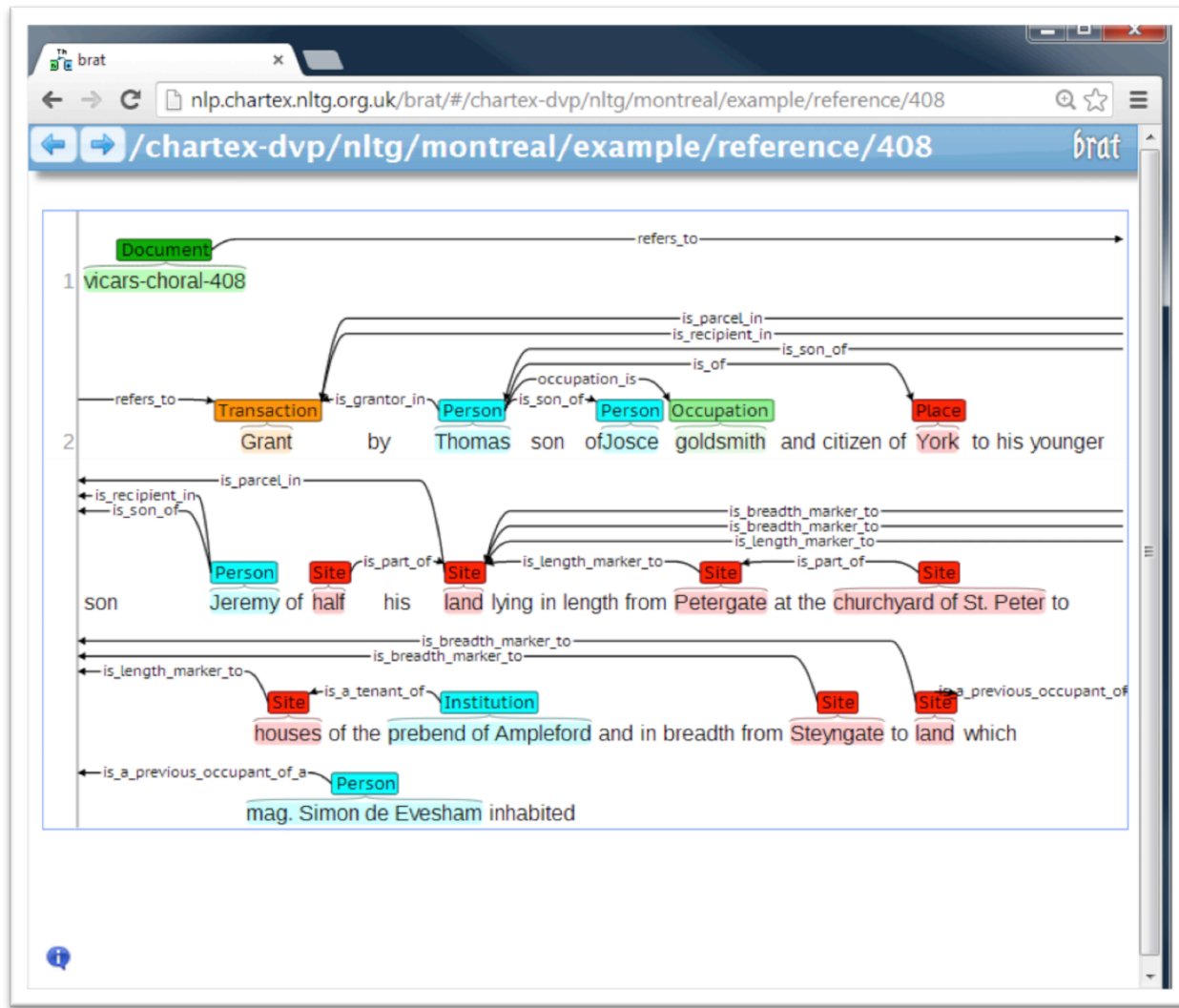
Web Semantics: Science, Services and Agents on the World Wide Web Volume 9, Issue 2 2011 113 - 127

<http://dx.doi.org/10.1016/j.websem.2011.02.001>

ChartEx development process

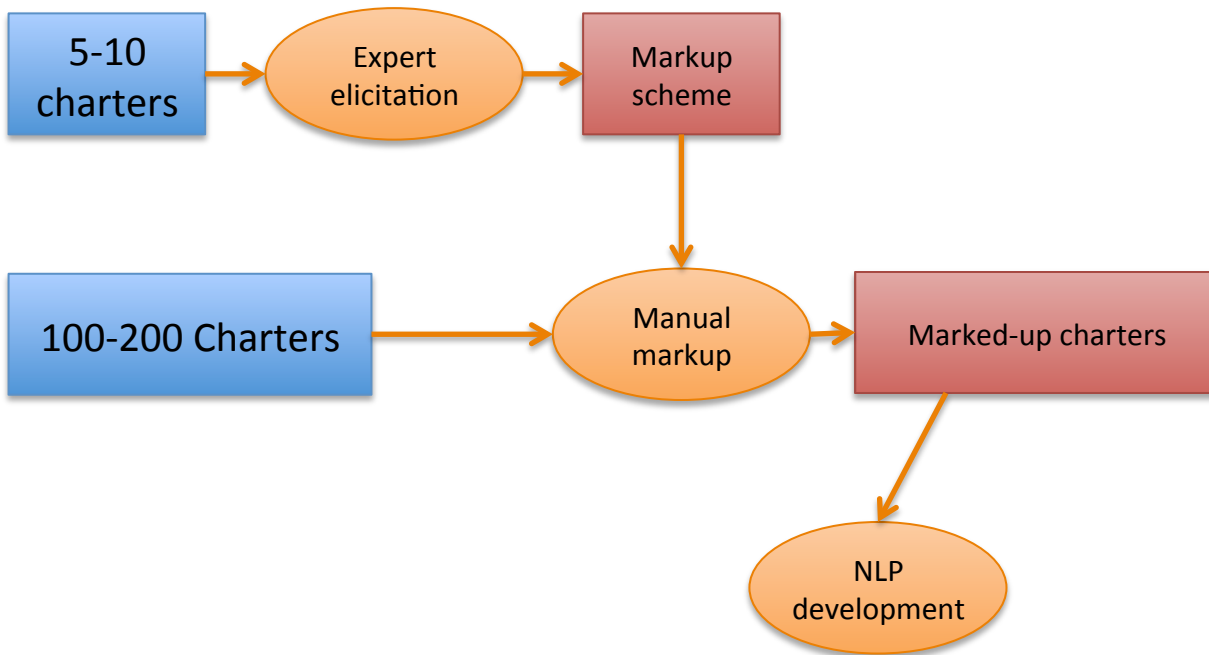


BRAT rapid annotation tool

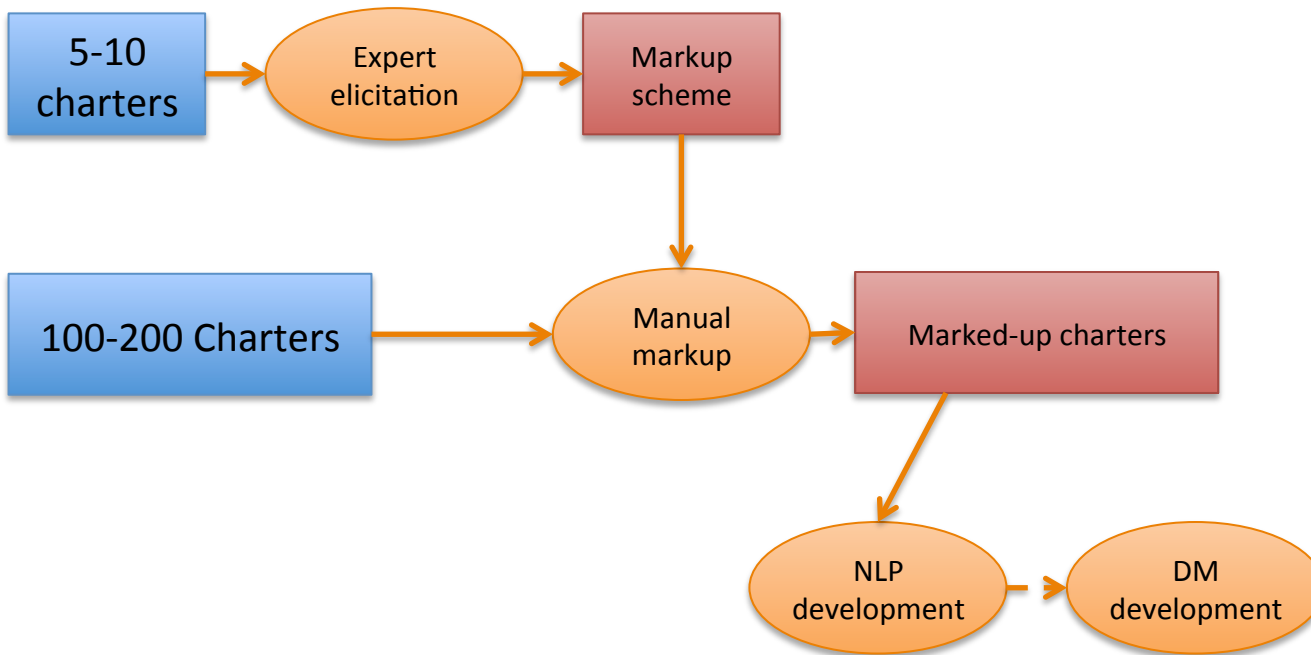


Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. (<http://brat.nlplab.org/>)

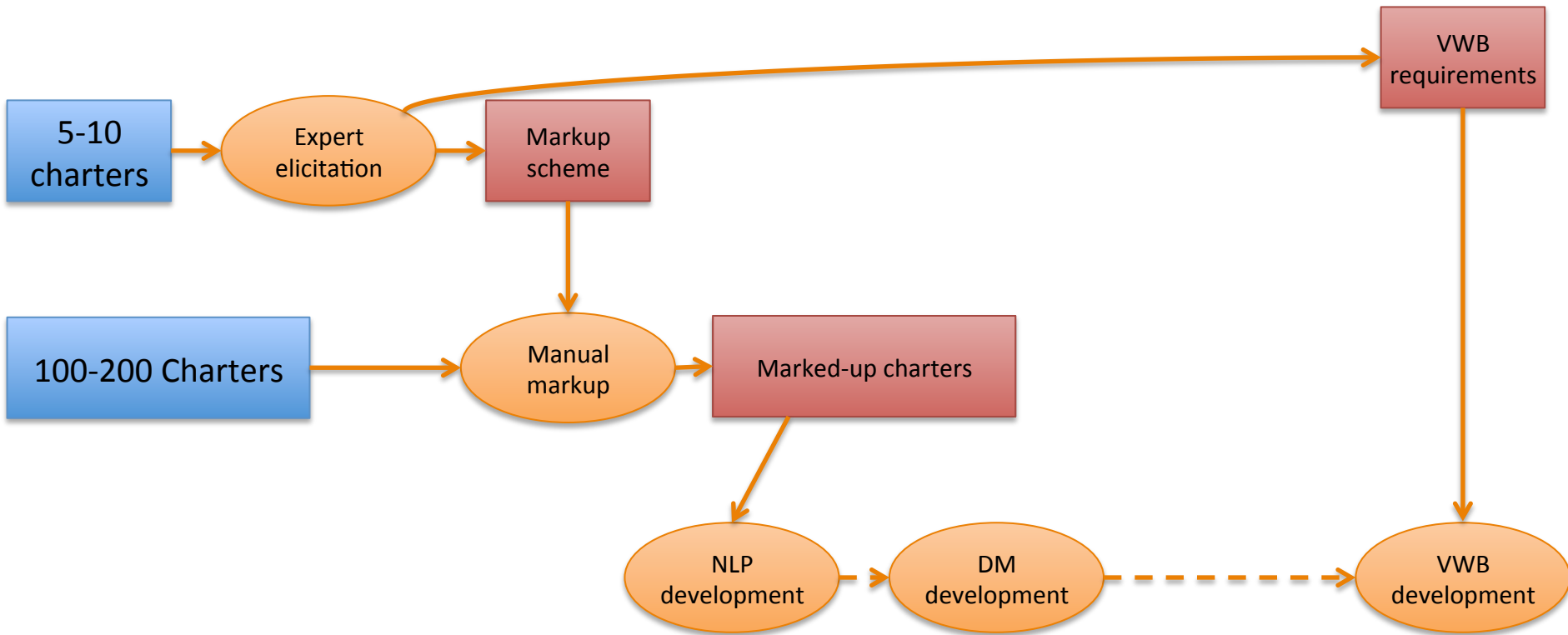
ChartEx development process



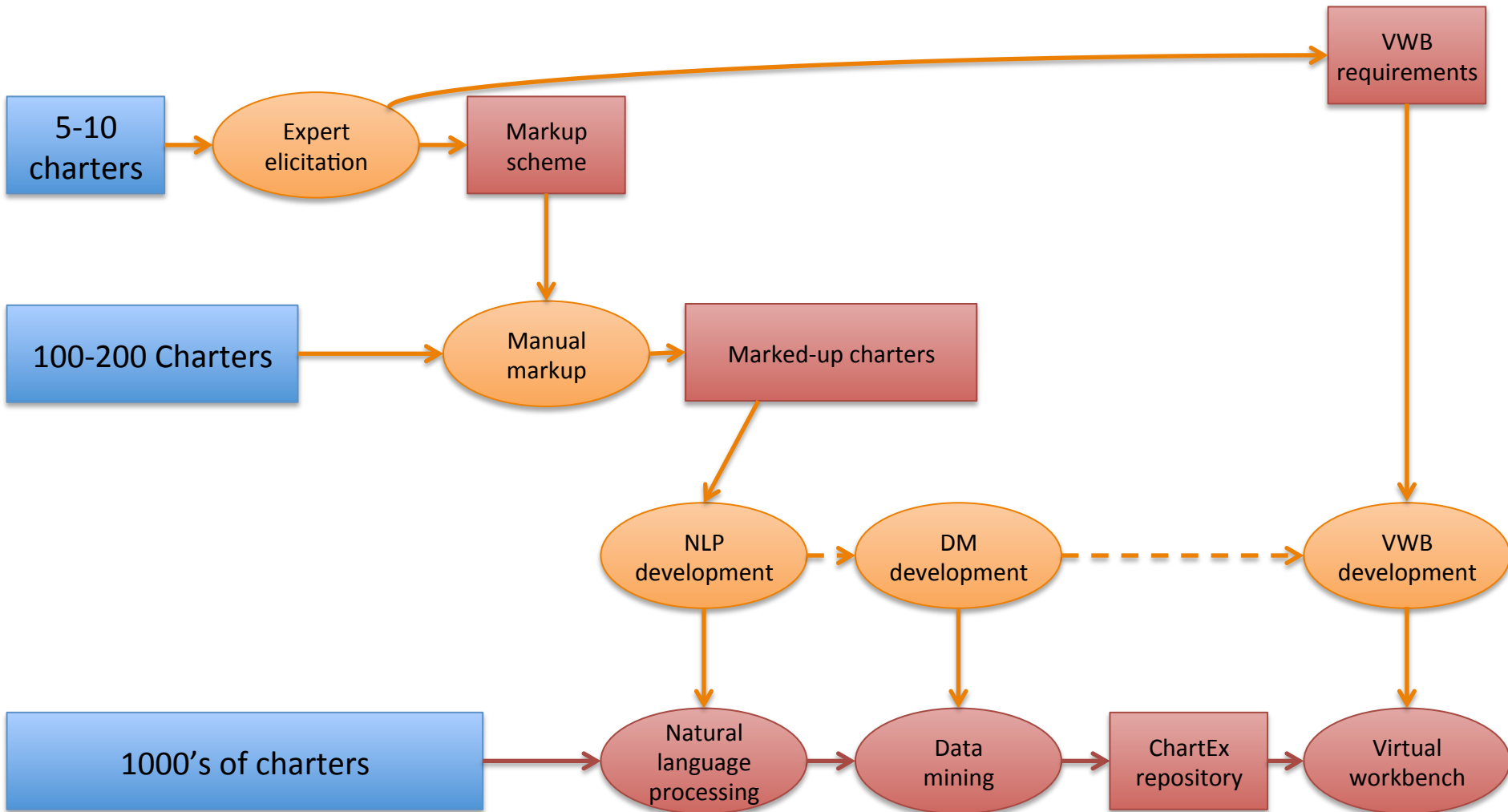
ChartEx development process



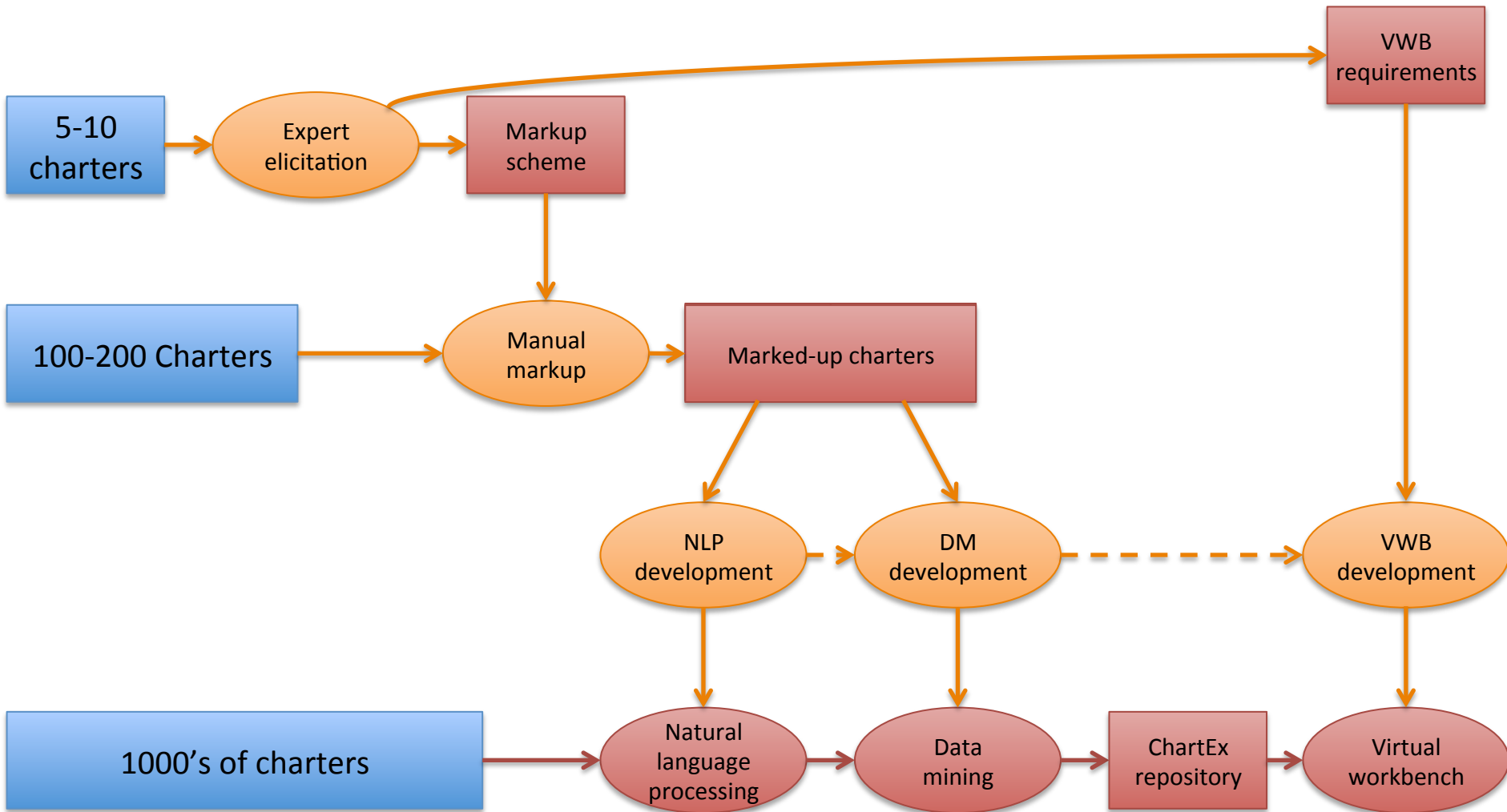
ChartEx development process



ChartEx development process



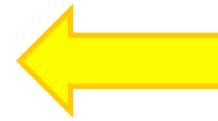
ChartEx development process



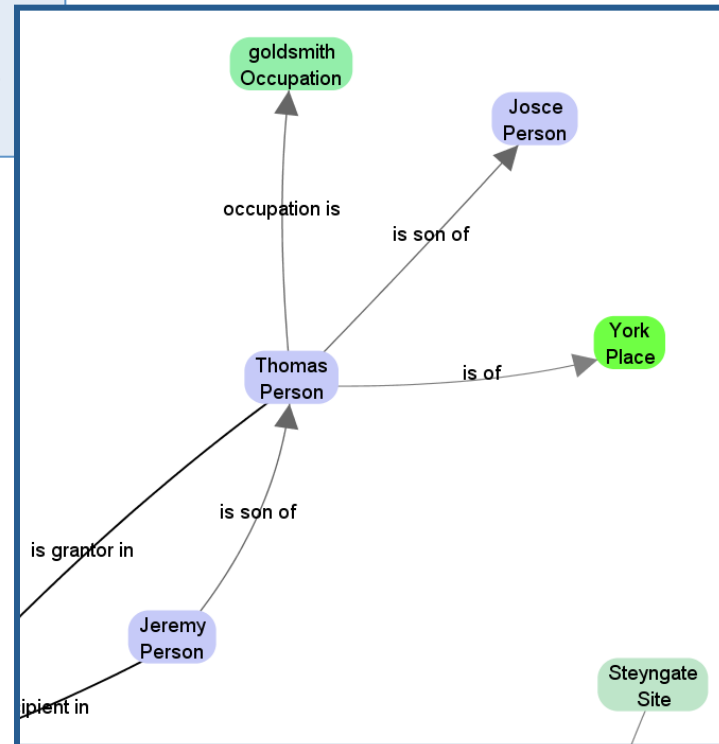
The NLP task

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.

Get from a text like this ...



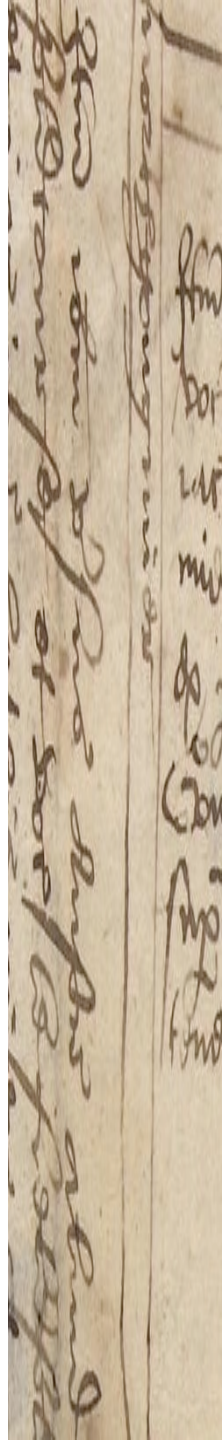
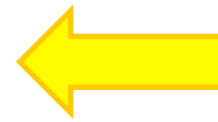
... to some
'semantic'
data like this



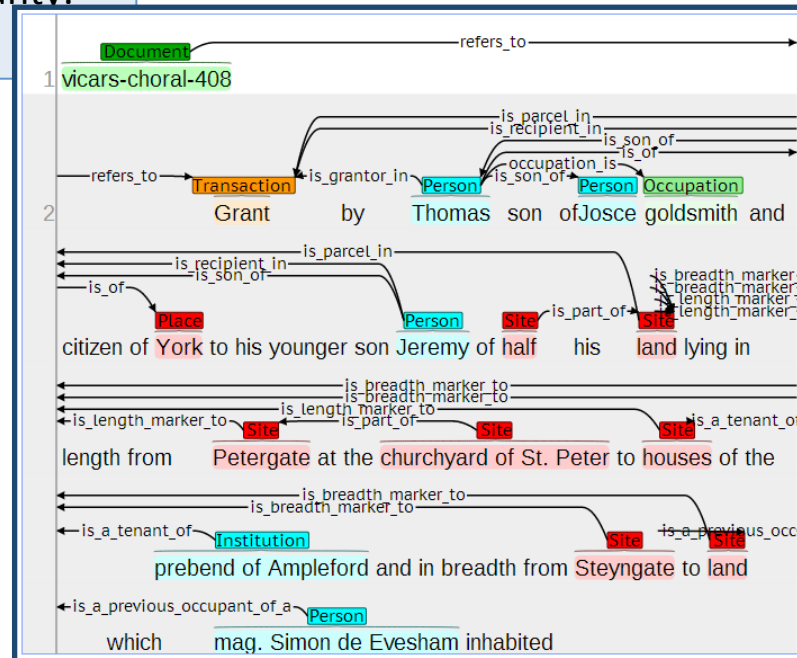
The NLP task

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.

Get from a text like this ...



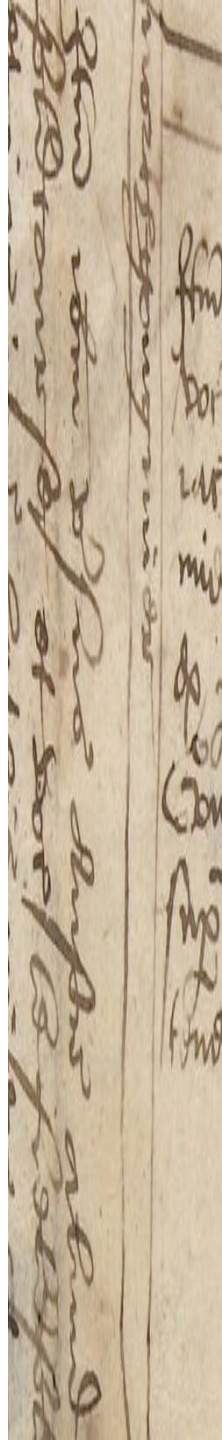
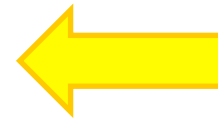
... which is not really very different from this (BRAT)



The NLP task

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.

Get from a text like this ...



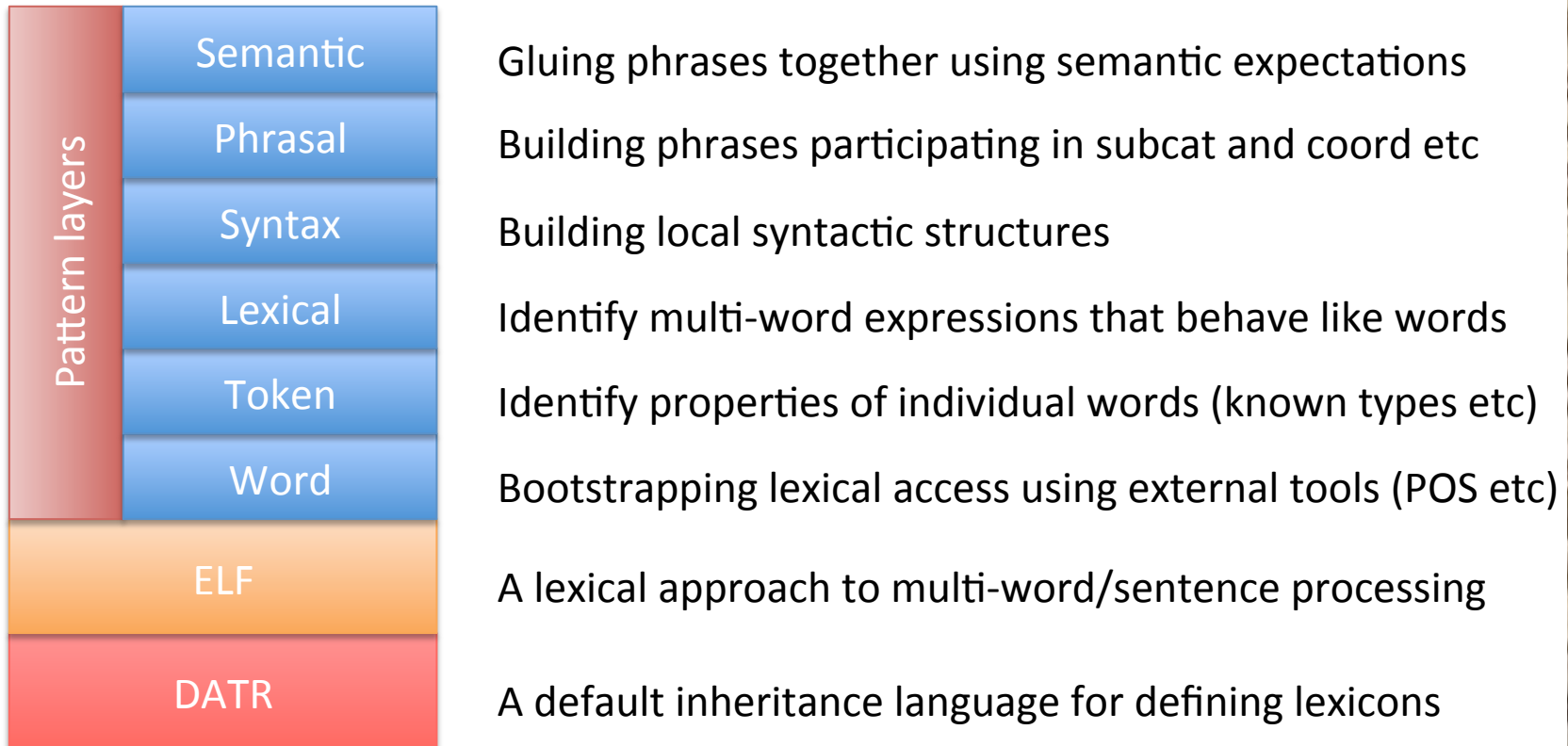
... although actually 'under the hood' it looks more like this



```
T1 Document 0 17 vicars-choral-408
T2 Transaction 18 23 Grant
T3 Person 27 33 Thomas
T4 Person 40 45 Josce
T5 Occupation 46 55 goldsmith
...
R5 refers_to Arg1:T1 Arg2:T2
R6 is_grantor_in Arg1:T3 Arg2:T2
R7 is_son_of Arg1:T3 Arg2:T4
R8 occupation_is Arg1:T3 Arg2:T5
```

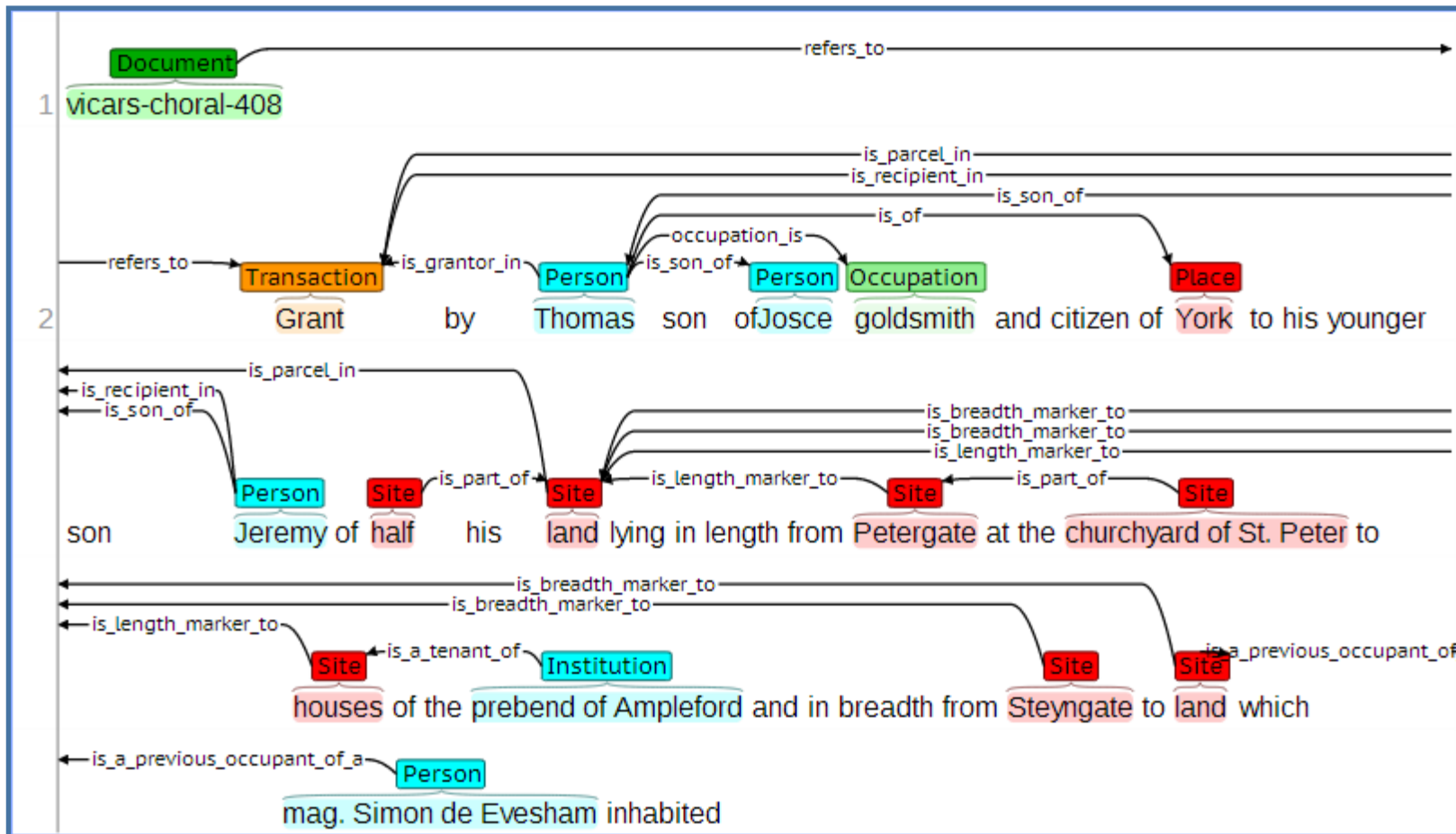
NLP approach

A (mostly) non-statistical, rule-based information extraction architecture.
Layered fst-like pattern matching defined in terms of default inheritance hierarchies.
Patterns (generalisations and exceptions) learnt from empirical data.
Flat semantics reminiscent of minimal recursion approaches.



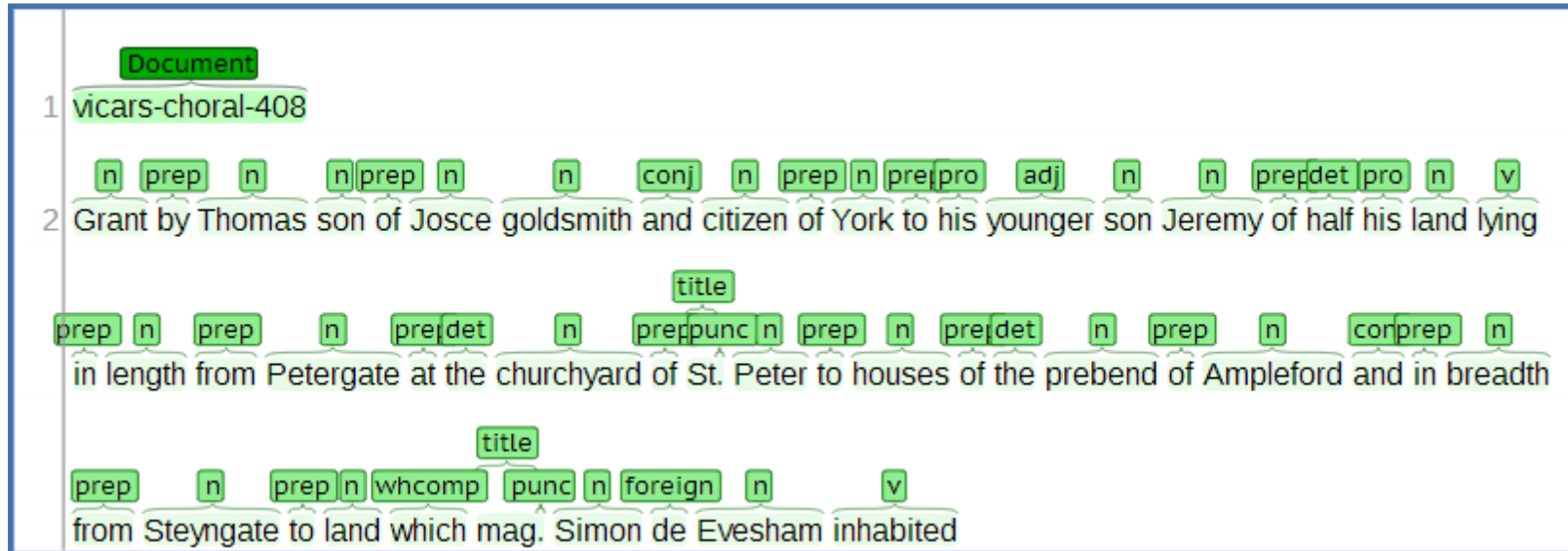
The manual annotation

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;



NLP – layered pattern matching

Word layer: basic information about individual words, collected ‘outside’ the main ChartEx NLP system (using external tools – part of speech tagger, stemmer, Soundex coding etc.)



NLP – layered pattern matching

Word layer: basic information about individual words, collected ‘outside’ the main ChartEx NLP system (using external tools – part of speech tagger, stemmer, Soundex coding etc.)

The image shows a screenshot of an NLP analysis interface. At the top, a green box labeled "Document" contains the text "vicars-choral-408". Below this, a line of text is displayed with various words and phrases highlighted in green boxes, representing different parts of speech or grammatical functions. A pop-up window is overlaid on the text, providing detailed information for the word "Josce".

1 vicars-choral-408

2 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 from Steyngate to land which

Property: word = josce
form = Josce
stem = josce
mform = count
mnum = sing
pos = n
casetype = capitalised
soundex = J200

NLP – layered pattern matching

Word layer: basic information about individual words, collected 'outside' the main ChartEx NLP system (using external tools – part of speech tagger, stemmer, Soundex coding etc.)

The image shows a screenshot of an NLP analysis interface. It displays a document titled "vicars-choral-408" and a snippet of text: "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 ... from Steyngate to land which mag. Simon de Evesham inhabited". The text is annotated with various parts of speech tags in green boxes, such as "n", "prep", "conj", "adj", "v", "title", "whcomp", "punc", and "foreign". A tooltip window is open over the word "lying", displaying its properties: "lying", "Property: word = lying", "form = lying", "stem = ly", "mform = part", "mtense = pres", "pos = v", "casetype = lowercase", and "soundex = I520".

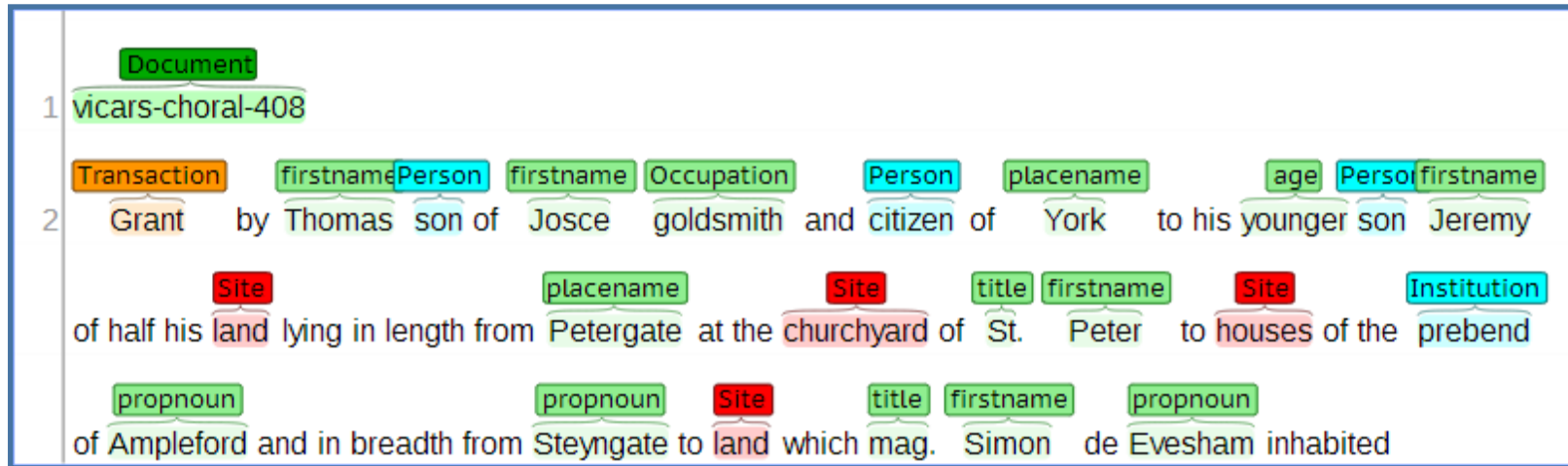
Document
1 vicars-choral-408

2 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 ... from Steyngate to land which mag. Simon de Evesham inhabited

Property: word = lying
form = lying
stem = ly
mform = part
mtense = pres
pos = v
casetype = lowercase
soundex = I520

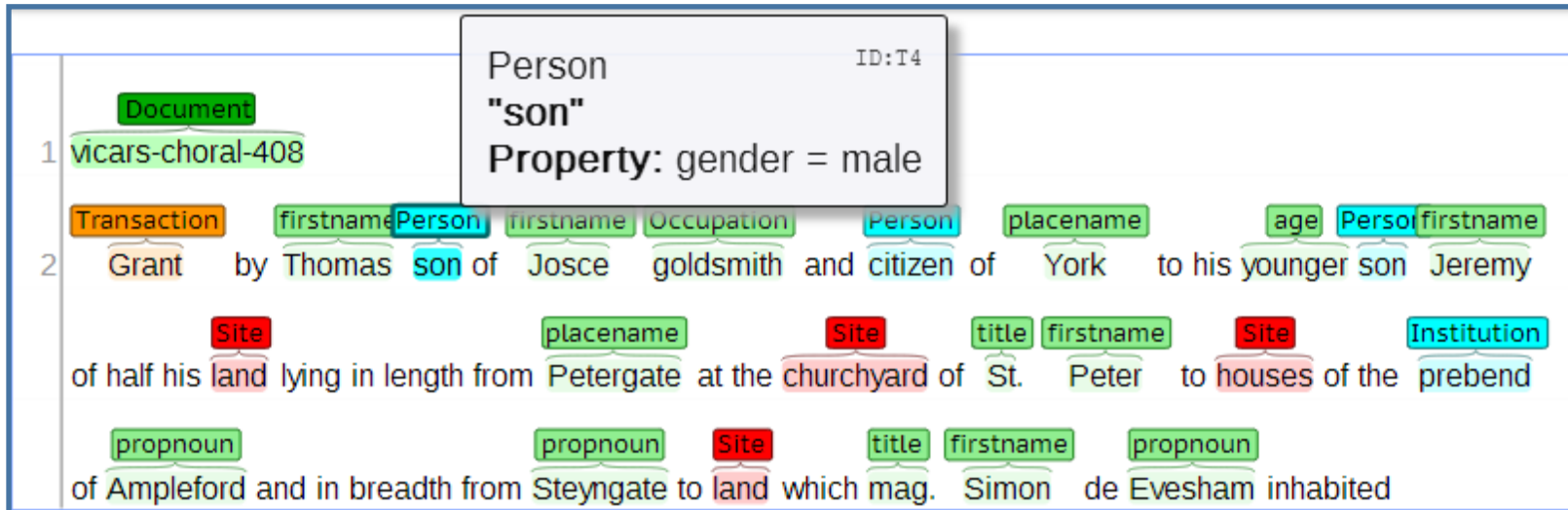
NLP – layered pattern matching

Token layer: identify semantic types and intrinsic properties (eg gender) of known individual words (not all of these types feature in the final output). Some types are inferred – eg unknown capitalised words are proper nouns.



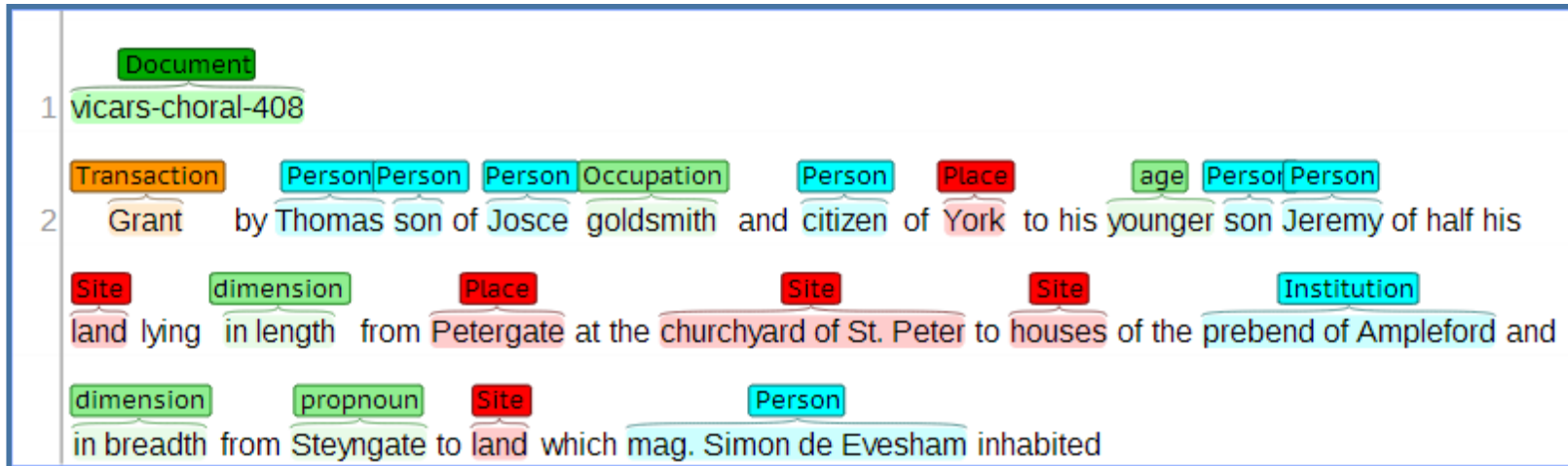
NLP – layered pattern matching

Token layer: identify semantic types and intrinsic properties (eg gender) of known individual words (not all of these types feature in the final output). Some types are inferred – eg unknown capitalised words are proper nouns.



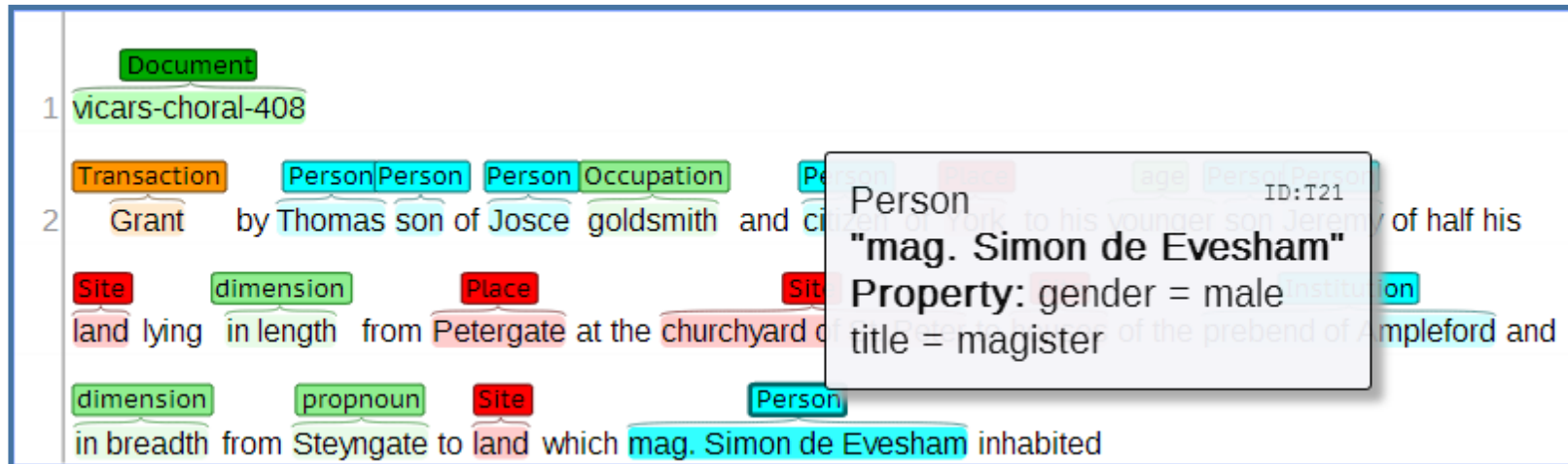
NLP – layered pattern matching

Lexical layer: identify simple lexical phrases – groups of tokens that act as individual units. Promote other individual tokens to lexical items (with lexical types).



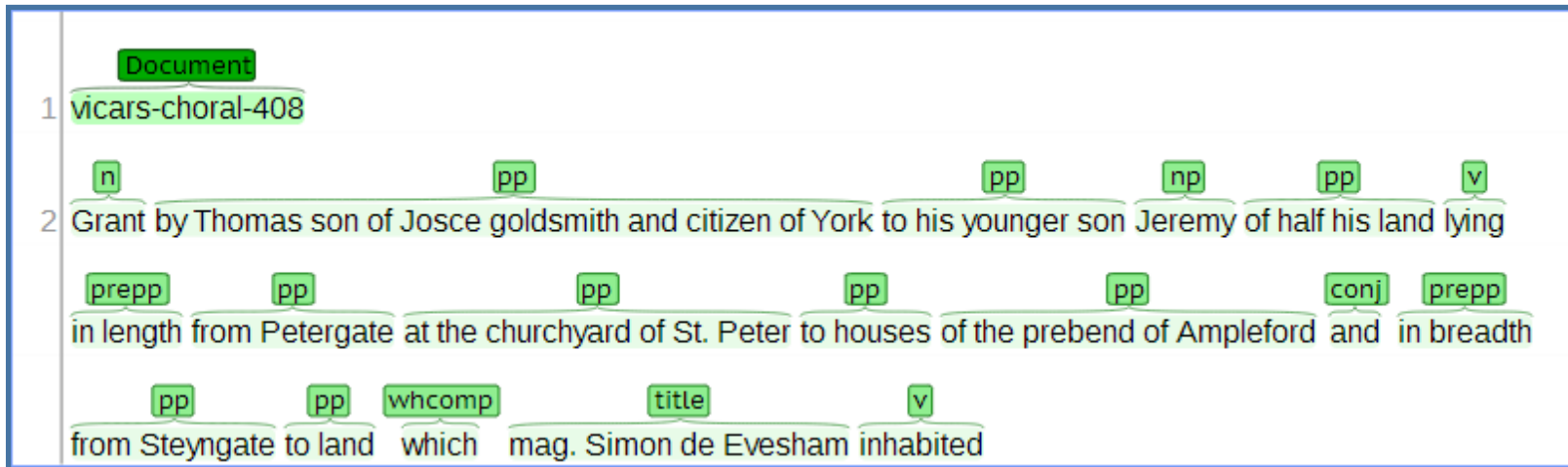
NLP – layered pattern matching

Lexical layer: identify simple lexical phrases – groups of tokens that act as individual units. Promote other individual tokens to lexical items (with lexical types).



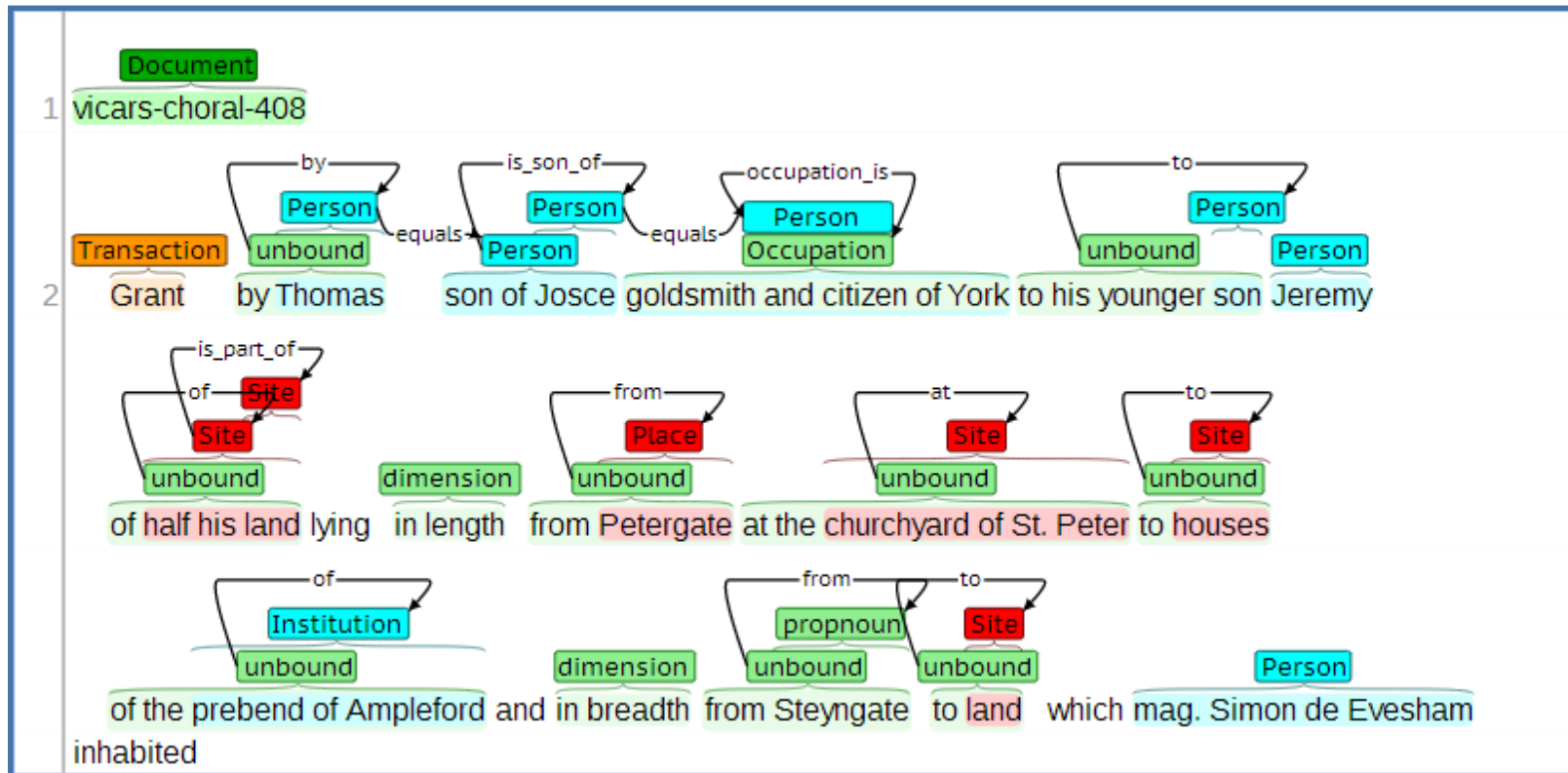
NLP – layered pattern matching

Syntax layer: build (local) syntactic structure to identify basic constituents of the sentence.



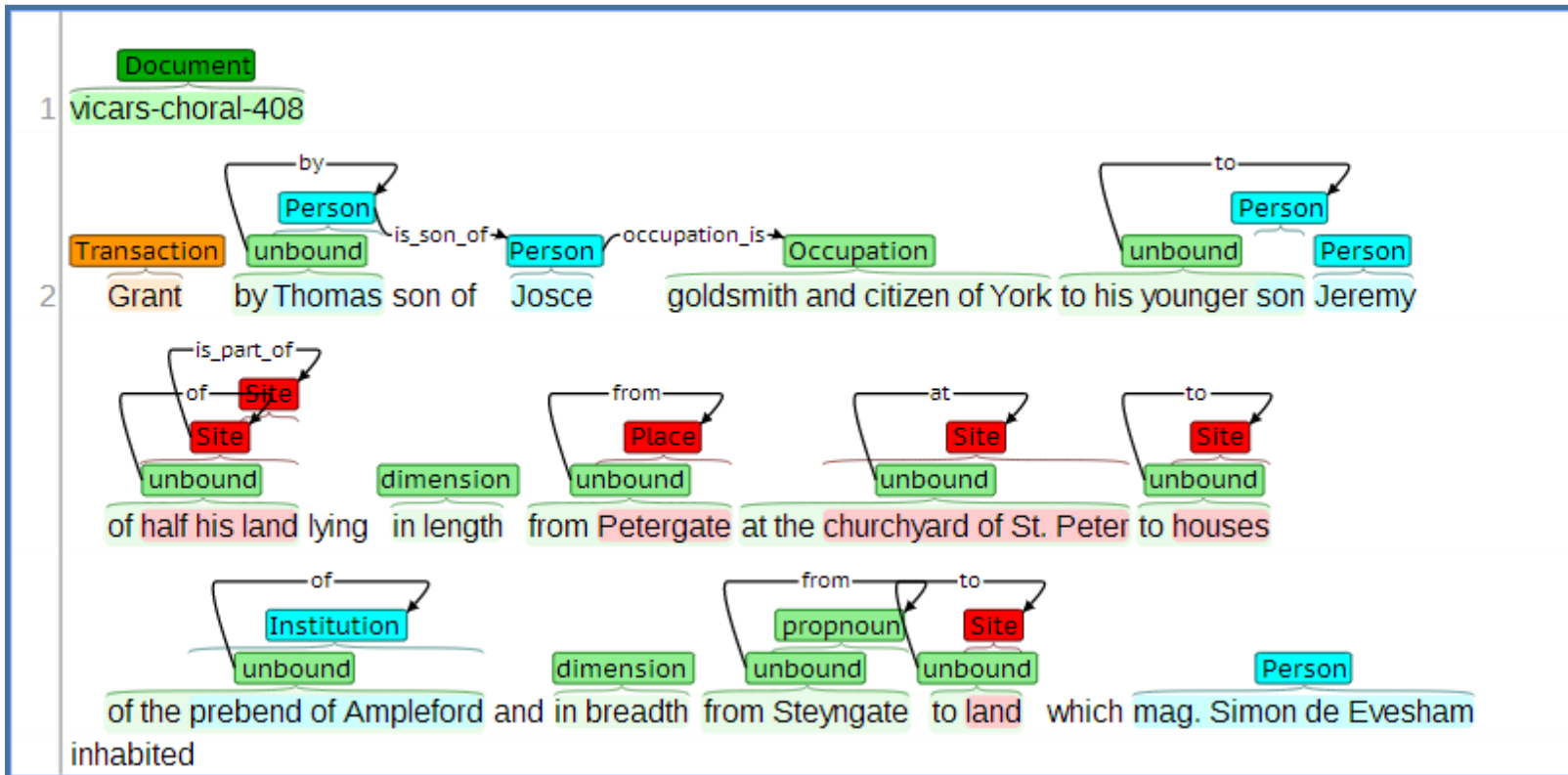
NLP – layered pattern matching

Phrasal layer: use part-of-speech tags to build lexical items into local syntactic/semantic structures. These have lots of **unbound** arguments – like jigsaw pieces waiting to be slotted together.



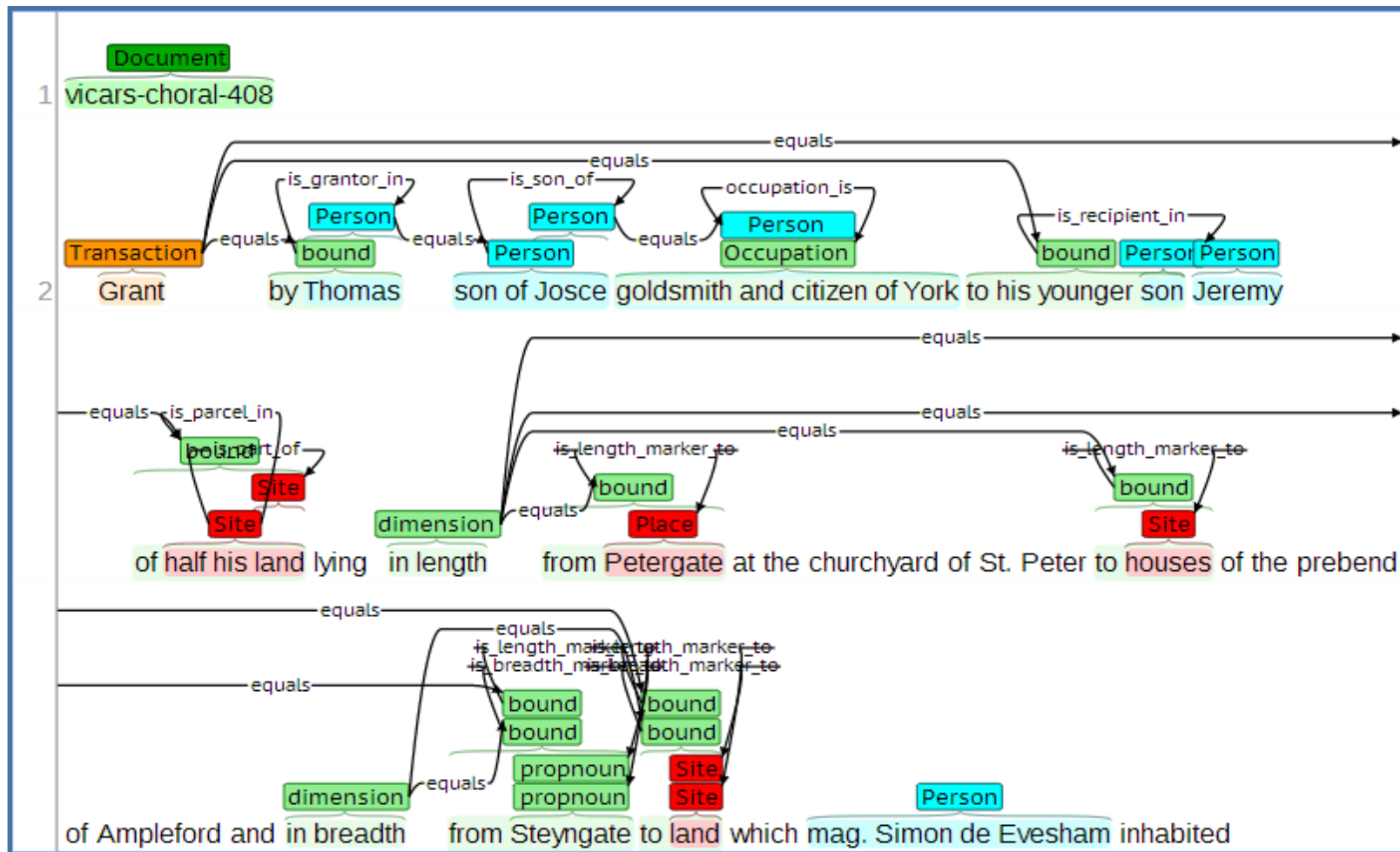
NLP – layered pattern matching

Phrasal layer: use part-of-speech tags to build lexical items into local syntactic/semantic structures – and then 'collapse' all the **equals** relations.



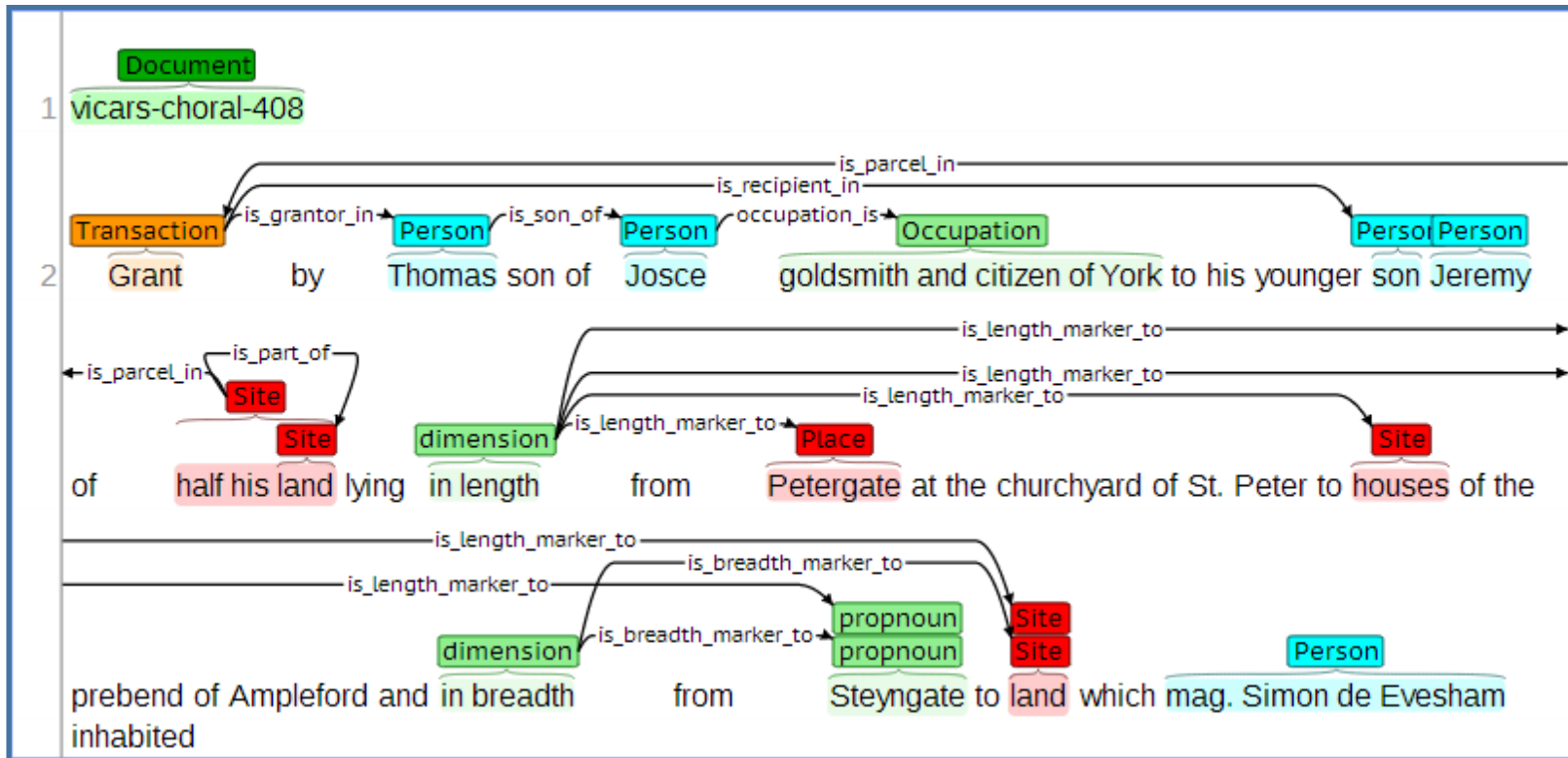
NLP – layered pattern matching

Semantic layer: build semantic relationships by gluing the pieces together. Again, first we use **equals** relations to capture the structure

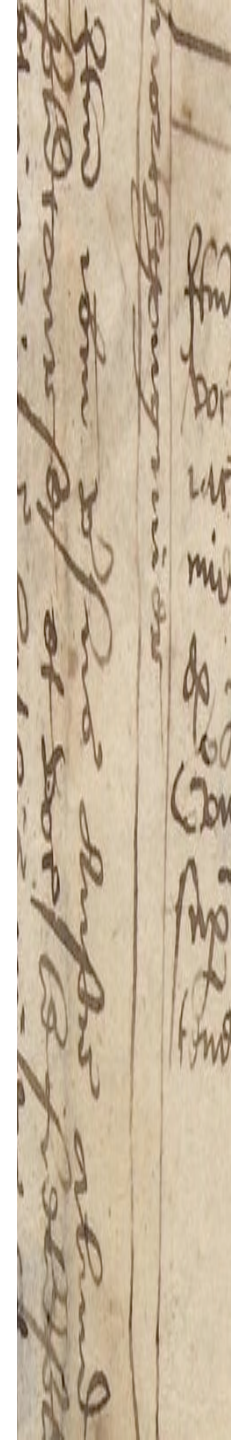
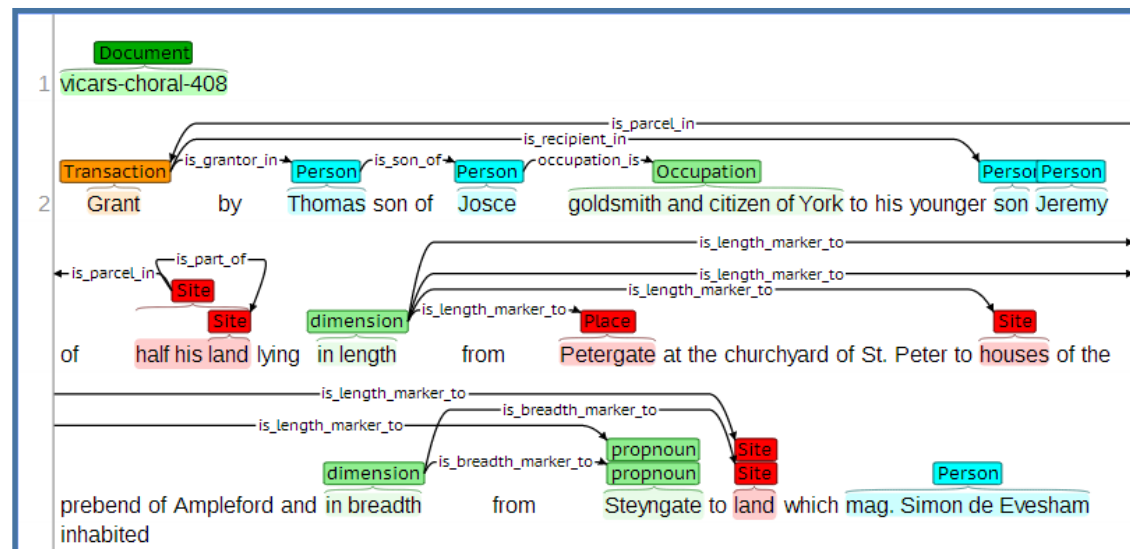
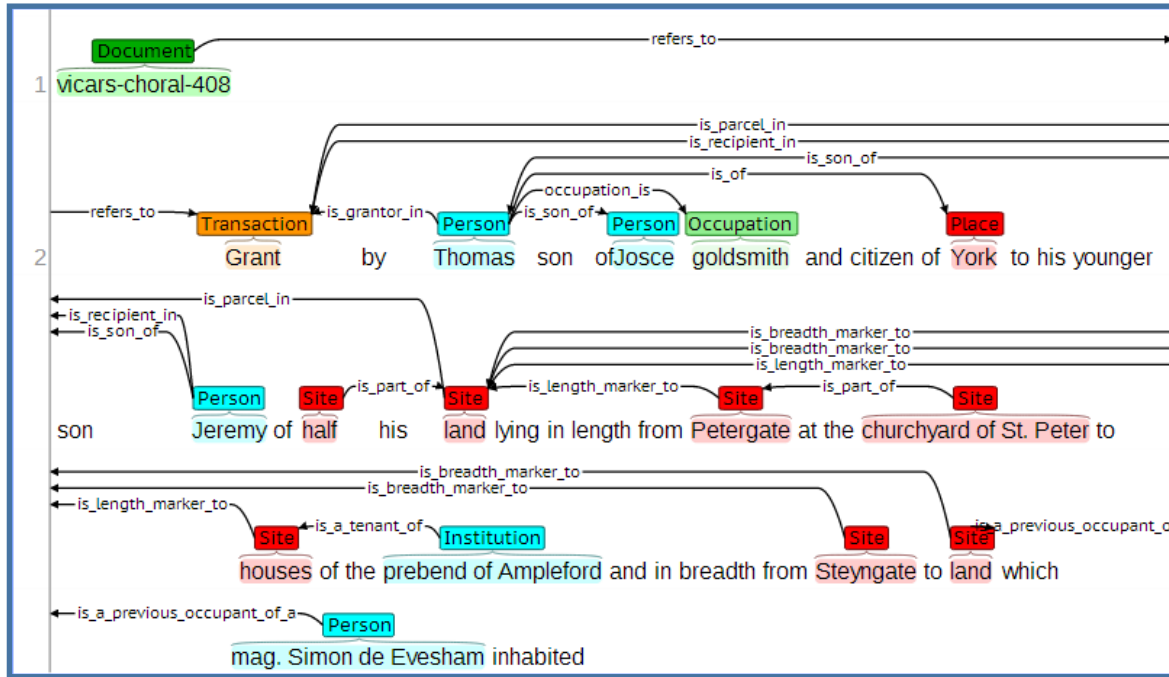


NLP – layered pattern matching

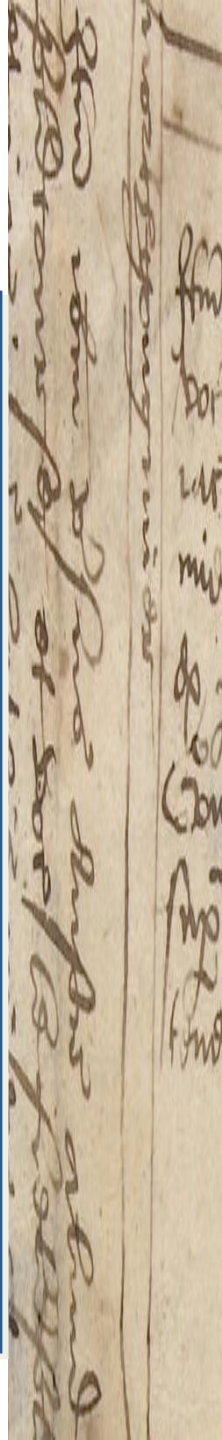
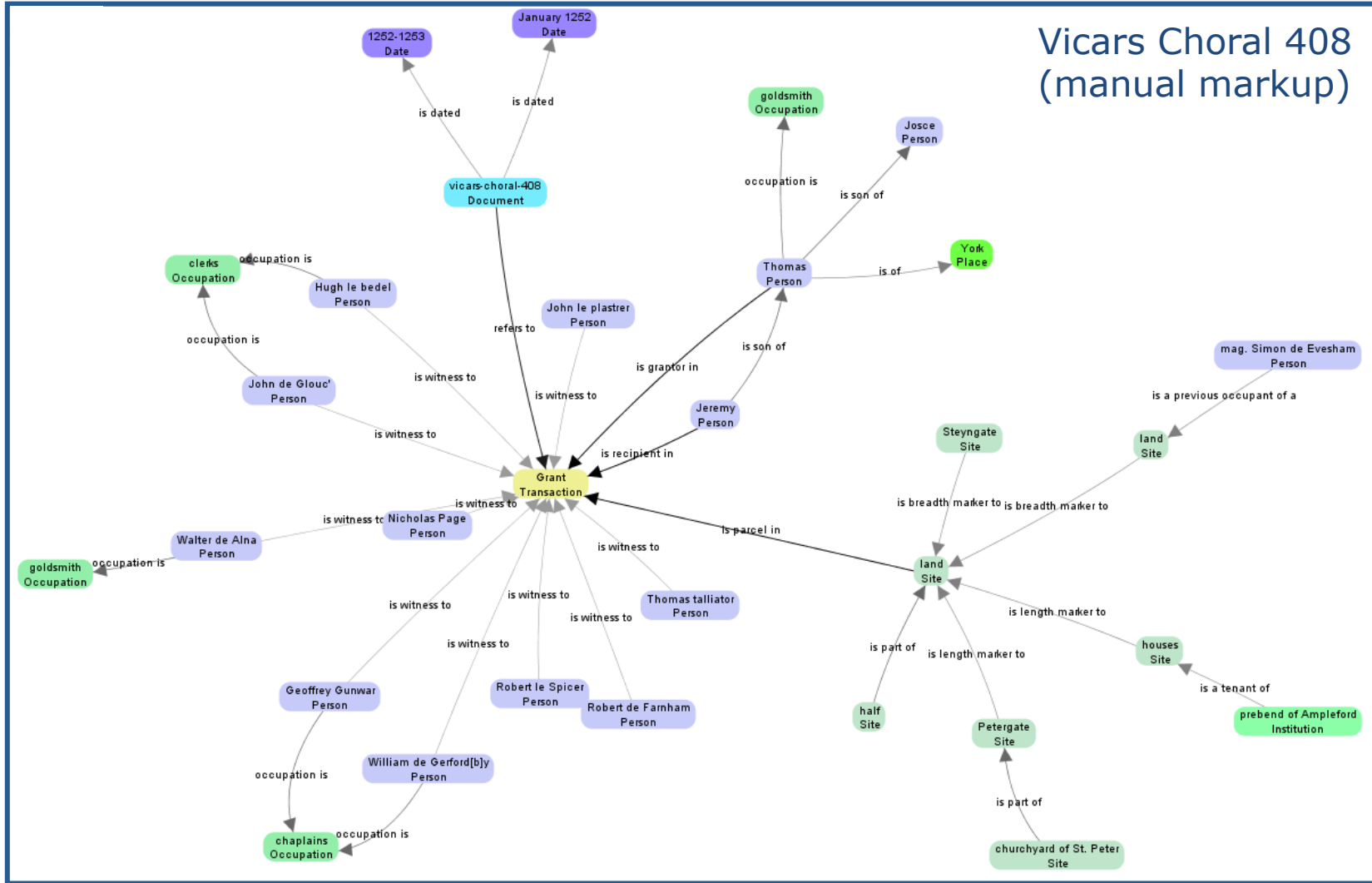
Semantic layer: build semantic relationships – and then we collapse them down



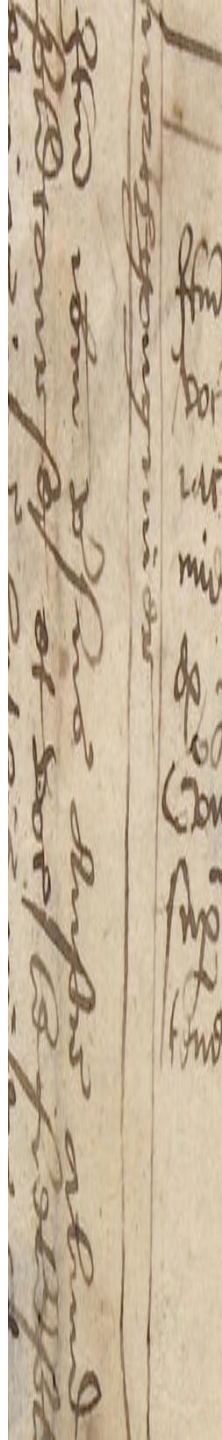
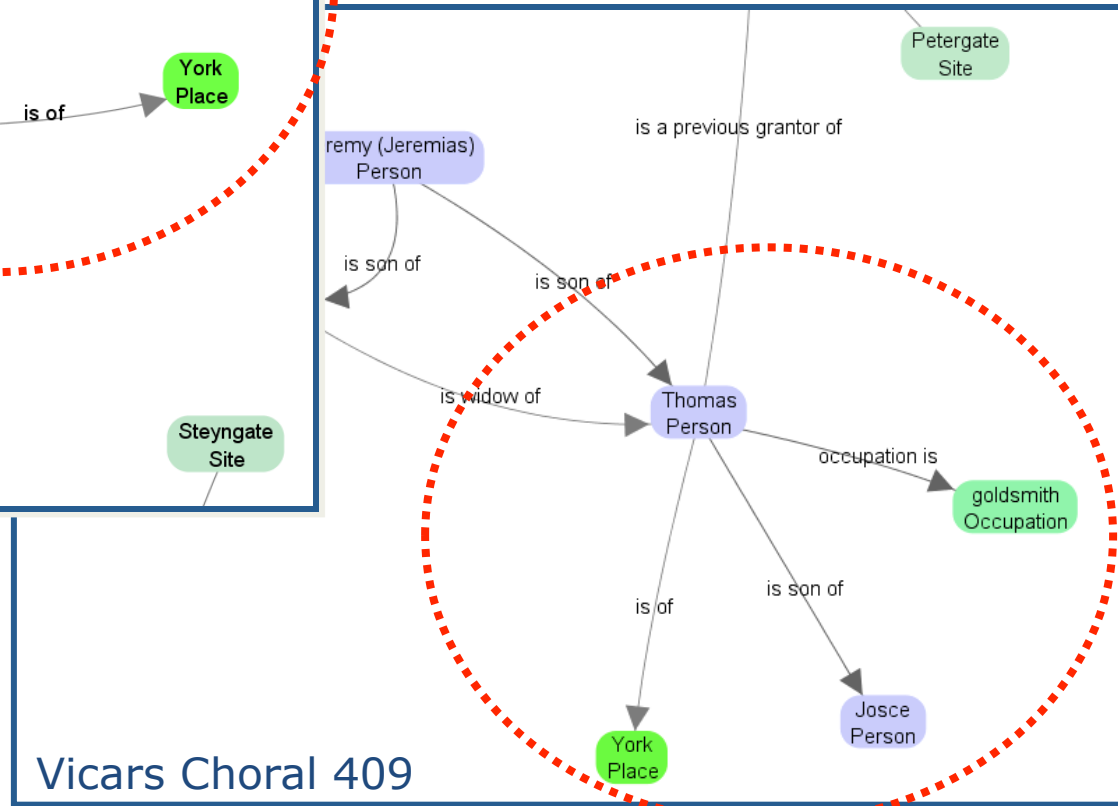
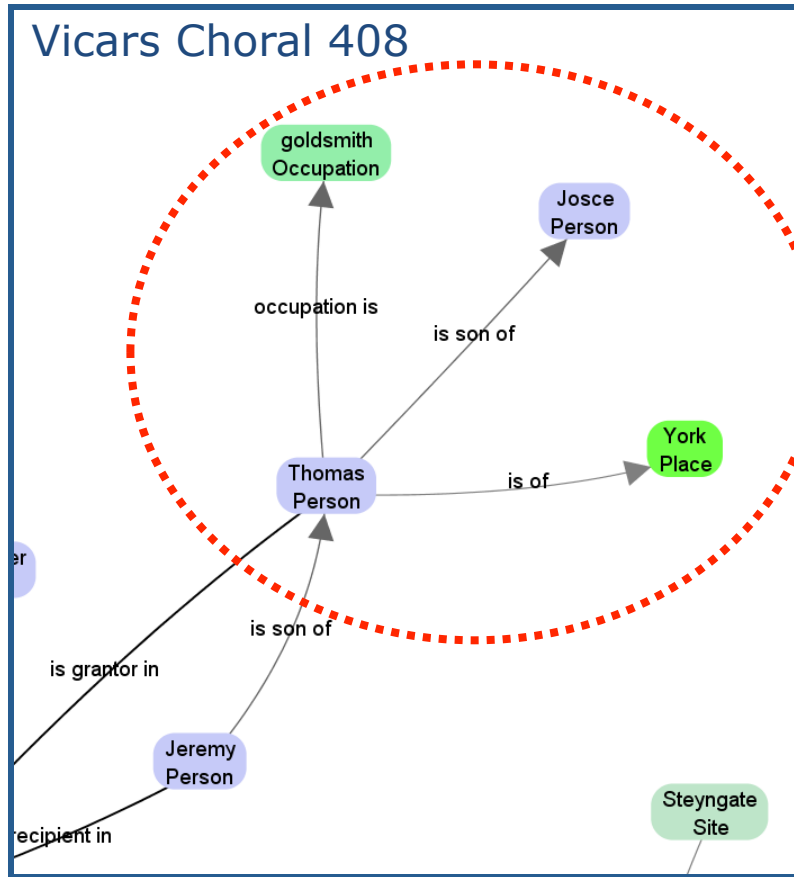
How did we do?



Data Mining – probabilistic record linkage



Matching information between charters



Probabilistic reasoning

“Thomas son of Josce, goldsmith”

- Statistics

$p(\text{Thomas}) = 0.12$ (common name)

$p(\text{Josce}) = 0.0015$ (uncommon name)

$p(\text{Goldsmith}) = 0.04$ (common profession)

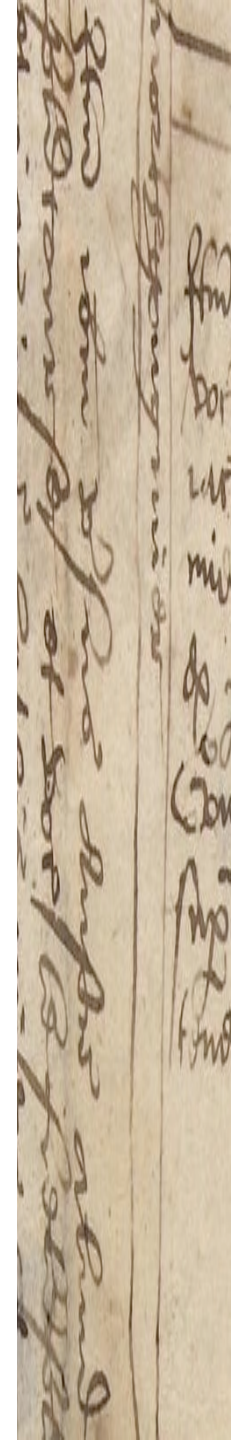
- Dating

vc-408 1252-1253

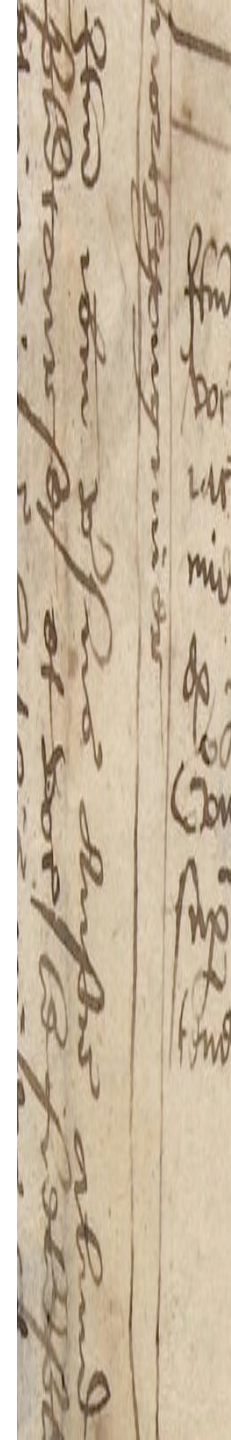
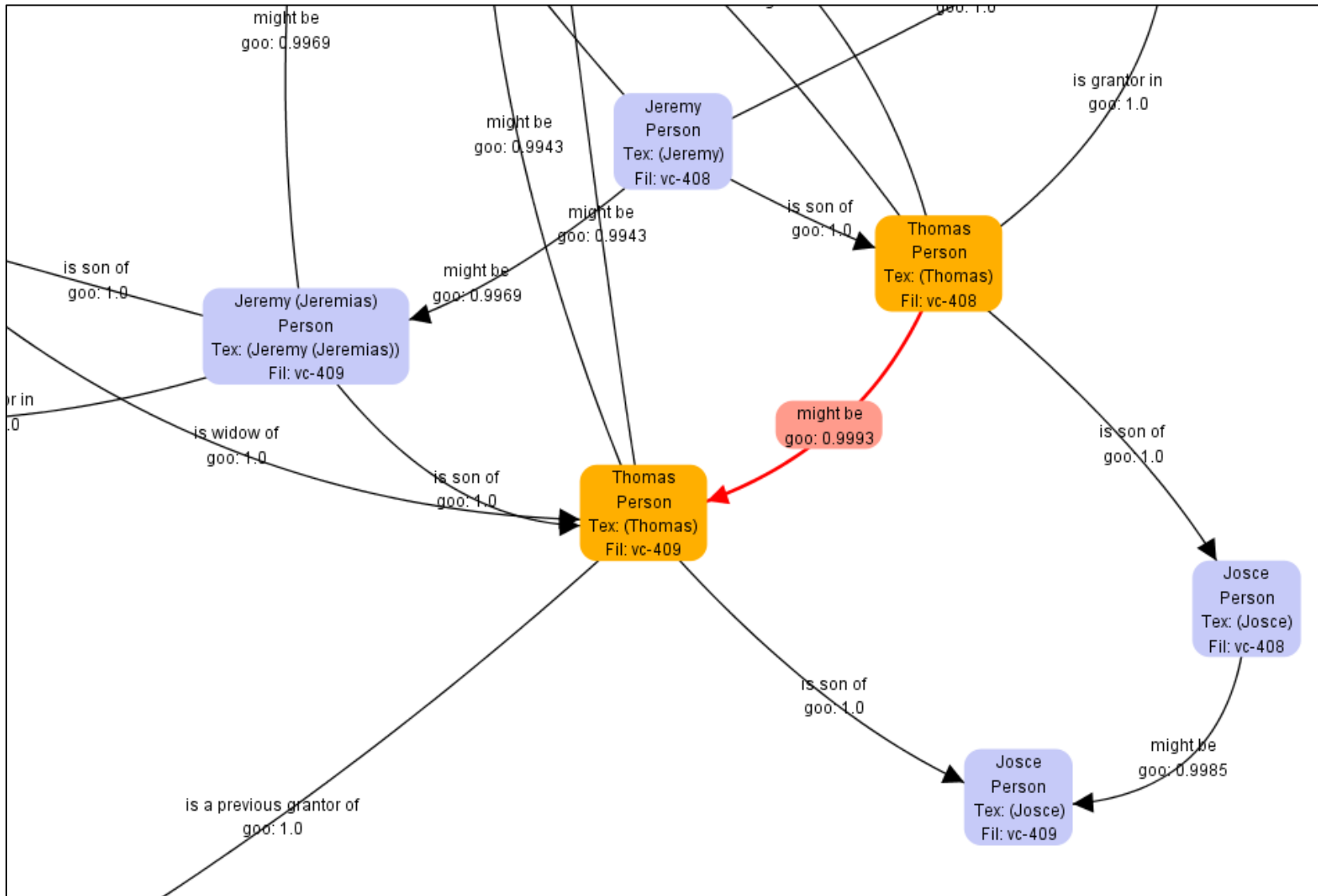
vc-409 1253-1261

- Final confidence

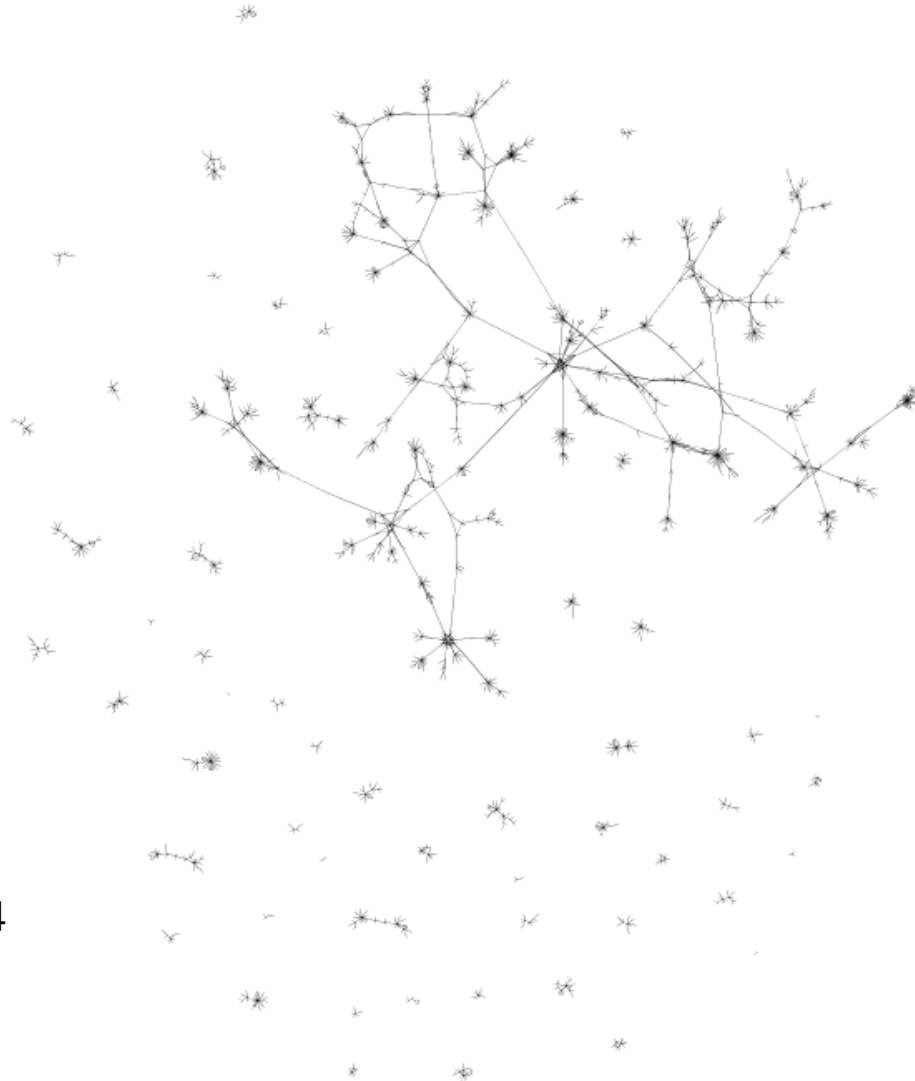
***conf* (Thomas 408, Thomas 409) = 0.9993**



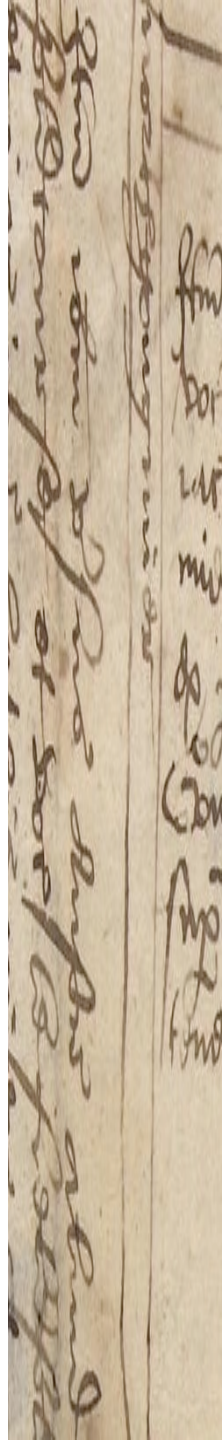
“Thomas son of Josce” matched



Cartularies as networks



Charters of the Vicars
Choral (selection of 124
charters)
confidence 0.99



The ChartEx Virtual Workbench

The screenshot displays the ChartEx Virtual Workbench interface, which is designed for analyzing historical documents. The interface is divided into several main sections:

- Search:** A search bar containing the term "goldsmith" and a "Search" button. Below it, there are options to filter results by collection, including "Vicars Choral: Goodramgate", "Vicars Choral: Petergate", and "Vicars Choral: General".
- Document Viewer:** Displays the text of a document titled "vicars-choral-411 (Vicars Choral: General)". The text is a historical record of a notification and a writ of novel disseisin, mentioning various individuals like John de Seley, Geoffrey Agulyun, and Alan Sampson.
- Entity Viewer:** Shows a list of entities related to the document, including "Person 805310320" and "Thomas (Person)". It also includes a "Possibly also..." section with other related entities.
- Transactions Visualisation:** A network diagram showing relationships between entities. Nodes represent individuals or places, and lines represent transactions or relationships between them.
- Table of Results:** A table listing search results for "goldsmith" across various collections. The table has columns for "Entity Extract", "Entity Type", "Document Name", and "Collection".

Entity Extract	Entity Type	Document Name	Collection
... de Roma, Radulph Caulte, Henry de capella, Galfrido de Otteley, Thomas the goldsmith, Simon longo, J. albo, Roaldo, et Reginaldo, chaplains and ...	Occupation	vicars-choral-388	Vicars Choral: General
... moine, Roger de St. Marzanet, chaplains, mag. John the physican, Thomas the goldsmith, Robert de St. Paul, Walter de Roma, Richard de Moserne, ...	Occupation	vicars-choral-404	Vicars Choral: General
... Alan Sampson, Gilbert de Fenton, William Blund, Ralph Furbur, Walter the goldsmith, Richard Moserne, Henry Blund, John the physican, Simon ...	Occupation	vicars-choral-406	Vicars Choral: General
Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his ...	Occupation	vicars-choral-408	Vicars Choral: General
... Robert de Farnham, Robert le Spicer, John le plastrer, Walter de Alna goldsmith, Nicholas Page, Thomas talliator, Hugh le bedel, John de ...	Occupation	vicars-choral-408	Vicars Choral: General
Grant by Mariot widow of Thomas son of Josce goldsmith of York and by Jeremy (Jeremias) son of Thomas and Mariot to ...	Occupation	vicars-choral-409	Vicars Choral: General
... mag. Jehn le Myre, William le Blund, Robert le Spicer, Walter de Alna goldsmith, John le plastrer, Richard Crapol, Richard le peyntur, ...	Occupation	vicars-choral-409	Vicars Choral: General
Notification by William, elder son and heir of Thomas of the churchyard goldsmith of York, that whereas a case was brought before the balliffs ...	Occupation	vicars-choral-410	Vicars Choral: General

<http://www.chartex.org/docs/Chartex-Workbench-Demonstration-VIDEO.mov>

The ChartEx Virtual Workbench

The screenshot displays the ChartEx Virtual Workbench interface, which is divided into several main sections:

- Search:** A search bar containing 'goldsmith' and a 'Search' button. Below it, there are options for 'Collections included in search:' and 'Advanced search (click to expand):'. The 'Document Results' and 'Entity Results' tabs are visible.
- Document Viewer:** Displays the document 'vicars-choral-408 (Vicars Choral: General)'. It shows the 'Document Text' with a snippet of Latin text: 'Grant by Thomas son of Josec goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchward of St. Peter to Rowens of the parsonage of Ampton and in breadth from Petergate to land which mag. Simon de Besham inhabited: paying Thome and his heirs id. or [a pair of] white gloves worth 3d. at Christmas, Warranty, Seal. Witnesses: Geoffrey Gunwar, William de Gerfordby, Chaplain, Robert de Farnham, Robert le Spicer, John le plasterer, Walter de Alna goldsmith, Nicholas Page, Thomas talliator, Hugh le bedel, John de Glouc, and others. January 1236. [1262/3]. SOURCE: VC 3/VI 386 (161 mm. x 137 mm.) ENDORSEMENT: Petergat, Donacio facta vicariis de domo que fait Thome aurfurhat; Simonis Evesham. SEAL: Silt. Hole in MS. NOTE: See 403.'
- Entity Viewer:** Displays the entity 'Person 805310320'. It shows 'Same as...' and 'Possibly also...' sections with lists of related entities like 'vicars-choral-408', 'Thomas (Person)', 'vicars-choral-136', etc.
- Transactions Visualisation:** A network diagram showing relationships between entities. Nodes include 'Thomas', 'Hugh le bedel', 'John de Glouc', 'Nicholas Page', 'Jard', 'Walter de Alna', 'is_witness_to', 'Grant', 'Document', 'is_grantor_in', 'is_recipient_in', 'Robert le Spicer', 'Robert de Farnham', 'William de Gerfordby', 'Geoffrey Gunwar', and 'Jeremy'. The diagram is set against a circular background.
- Person Visualisation:** A network diagram showing relationships between entities. Nodes include 'Thomas', 'vicars-choral-412', 'vicars-choral-411', 'vicars-choral-408', 'Same As...', 'Person 805310320', 'Possibly Also...', and 'vicars-choral-409'.

Conclusion

- ChartEx has been a very fruitful collaboration between historians and computer scientists
- ChartEx is not quite finished – still working on integration, evaluation and Latin.
- The project has achieved a great deal on quite modest budgets and timescales (would be an excellent first two years of a three year project ...)
- Aim to deliver as much of this as we can to the wider community
- Pursuing opportunities for follow-on research (funding, partners etc.)

