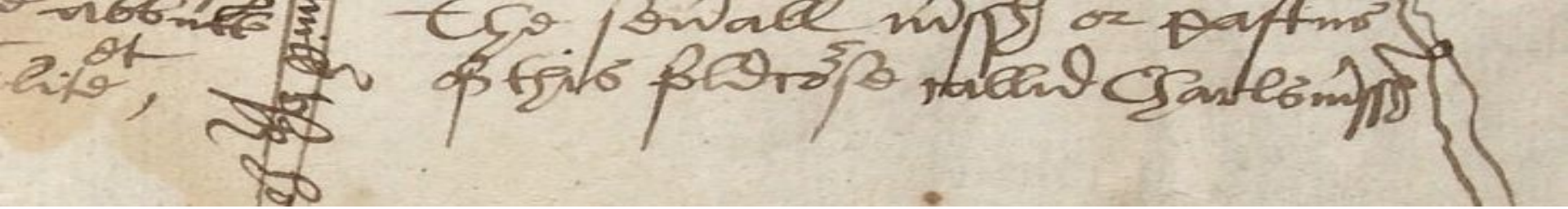


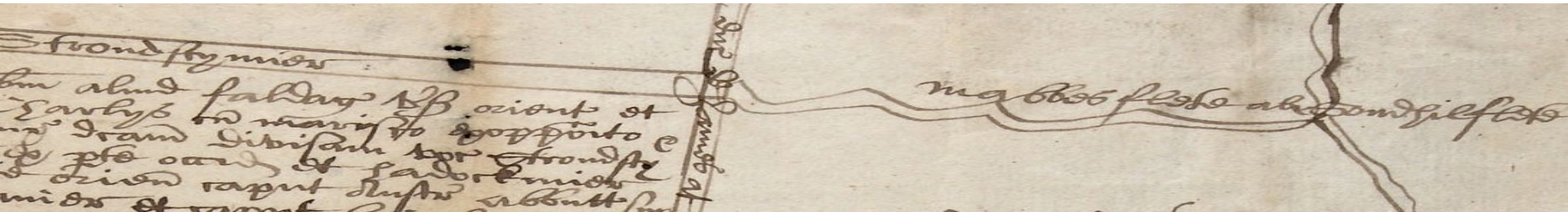
# Exploring medieval charters: the ChartEx Project


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')

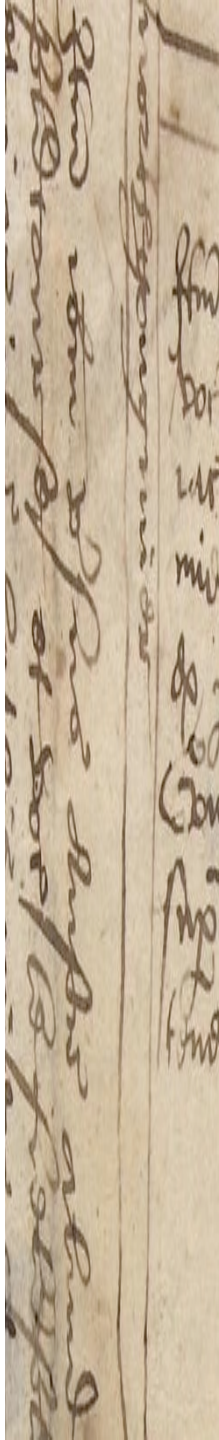
- Charters are among the oldest written records of 'everyday' European life and society – going right back to 1100's
- Early charters were all in Latin – also in English from around 1300's
- Thousands of charters survive, handwritten documents on delicate scrolls
- Charters contain lots of information 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.*

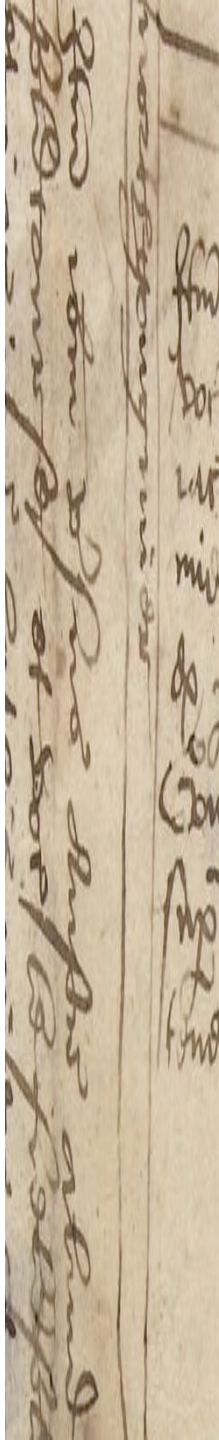


# Example Charter

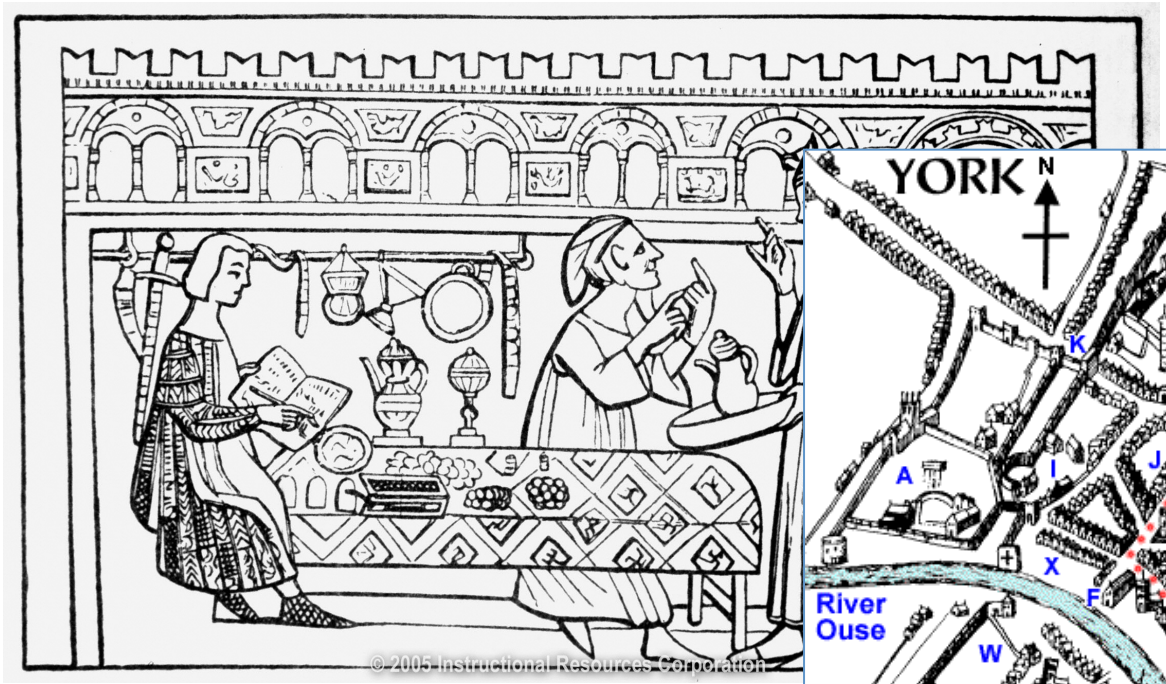
(from the Vicars Choral collection, University of York)

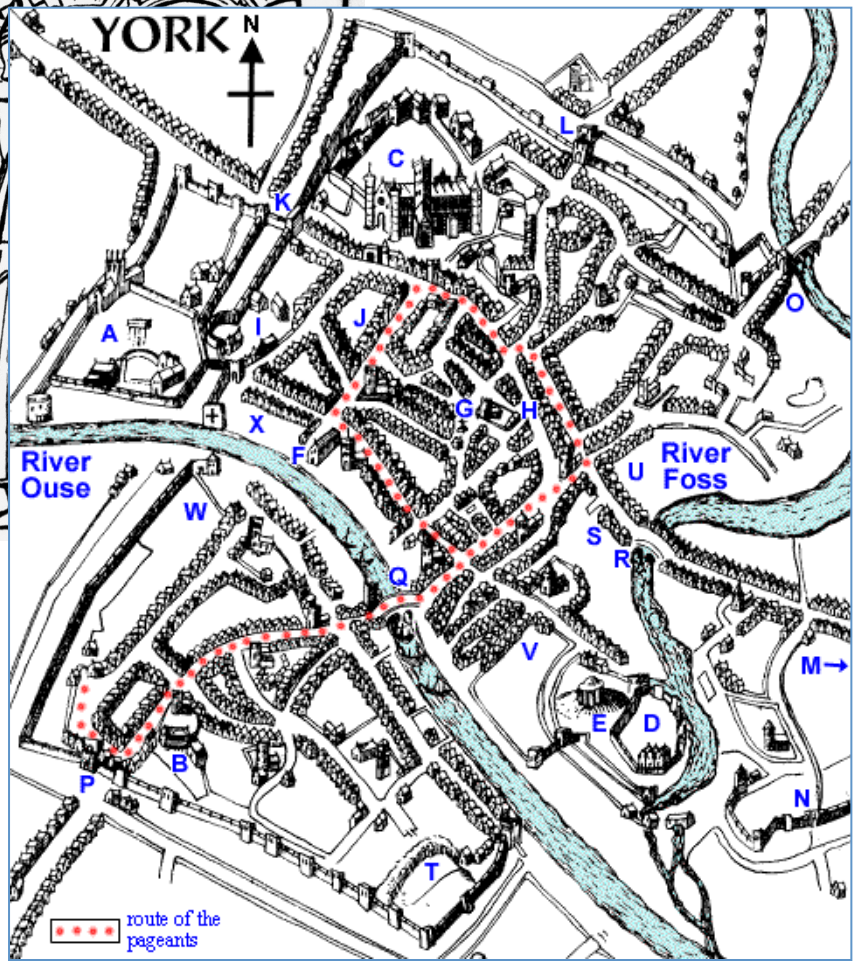
408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and 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

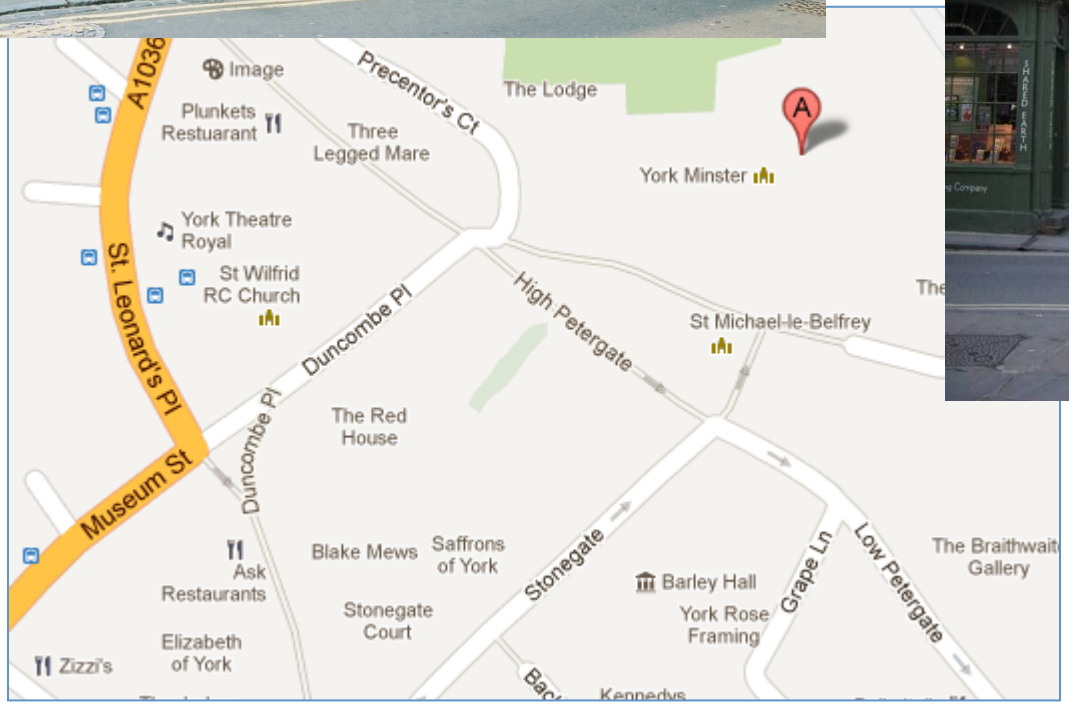


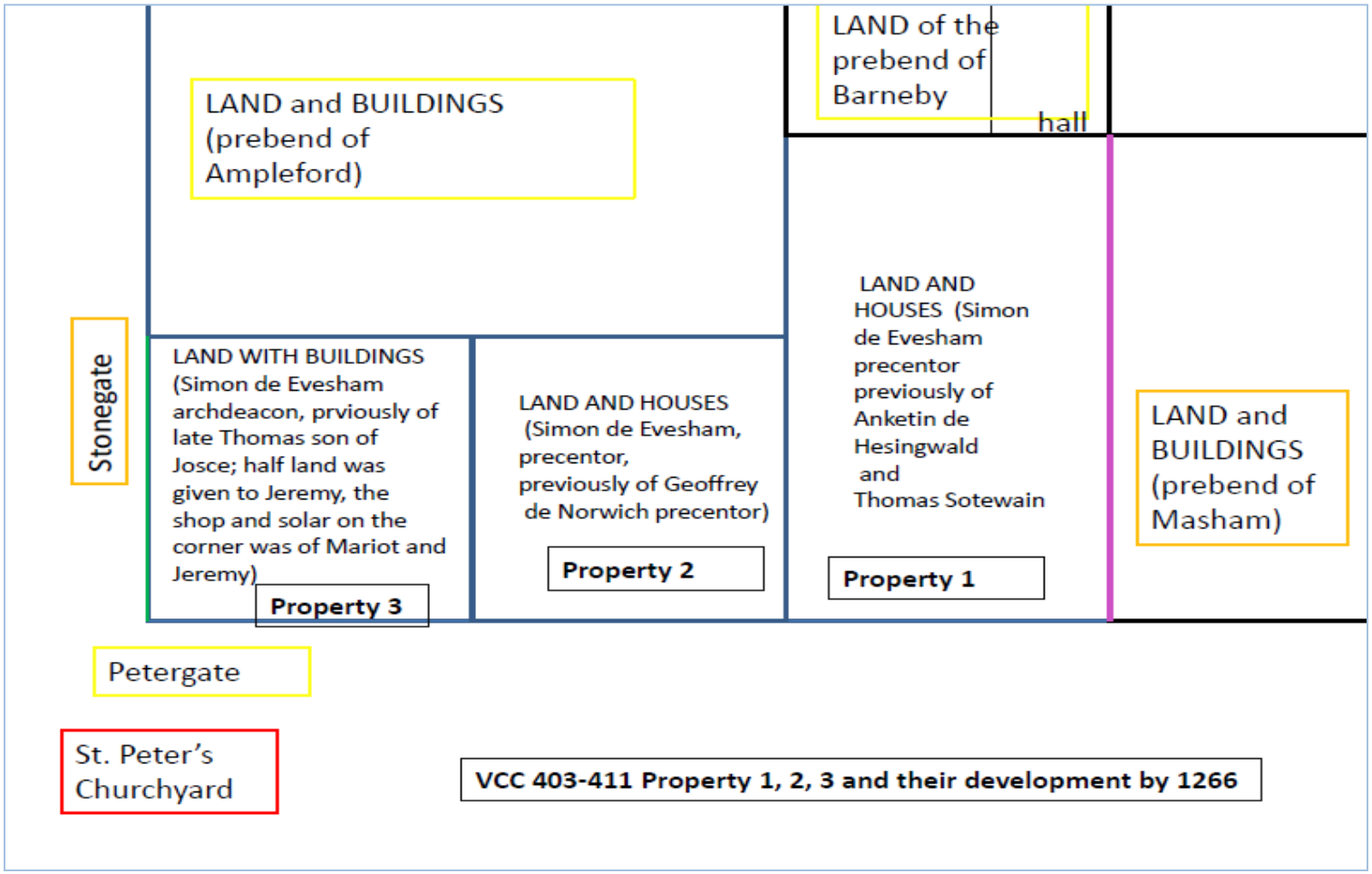

Medieval Petergate (York)

# Petergate today





# Conceptual plan of logical relationships




N

# 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

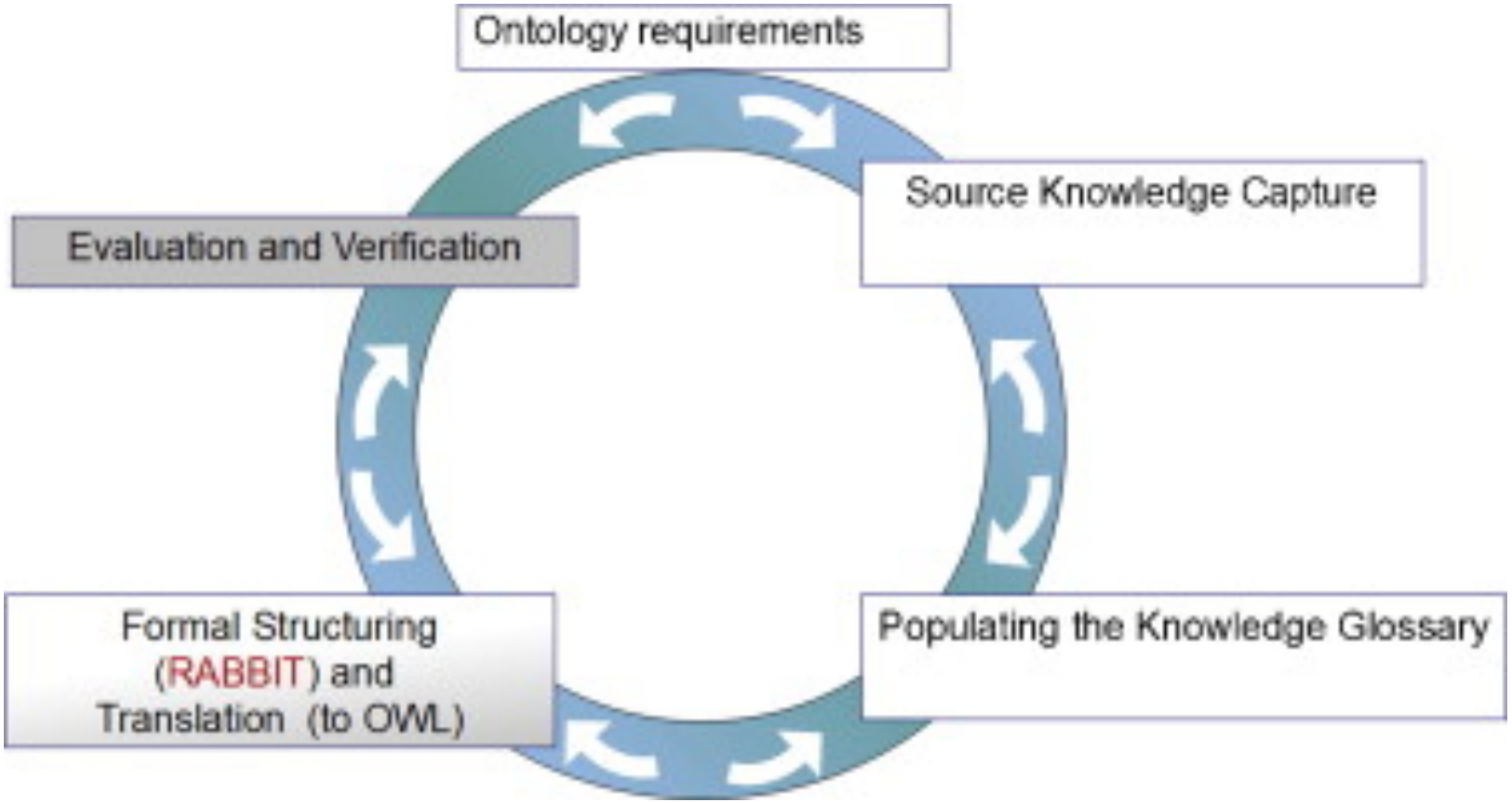


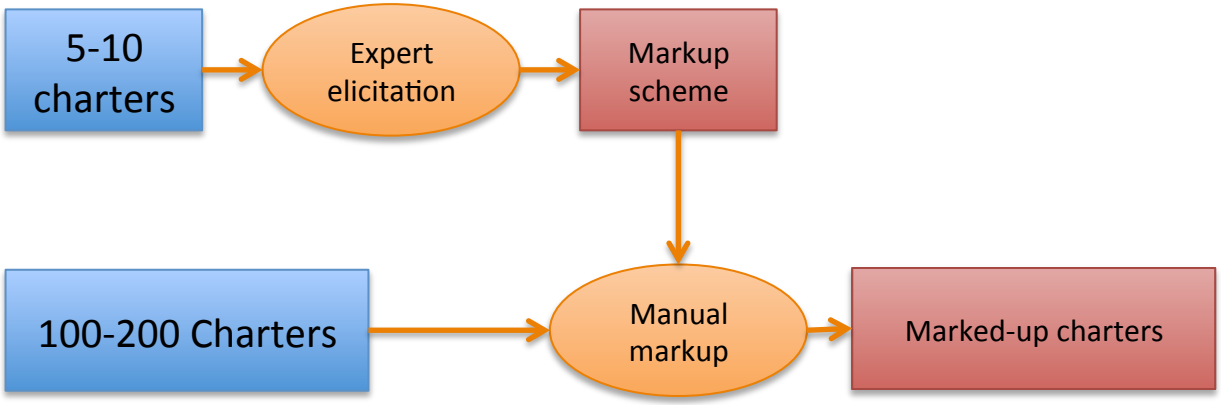

Fig. 1 The phases of the Kanga Methodology. The white boxes indicate the phases performed by domain experts. The formal structuring is done by domain experts using Rabbit, while the translation to OWL is ...

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

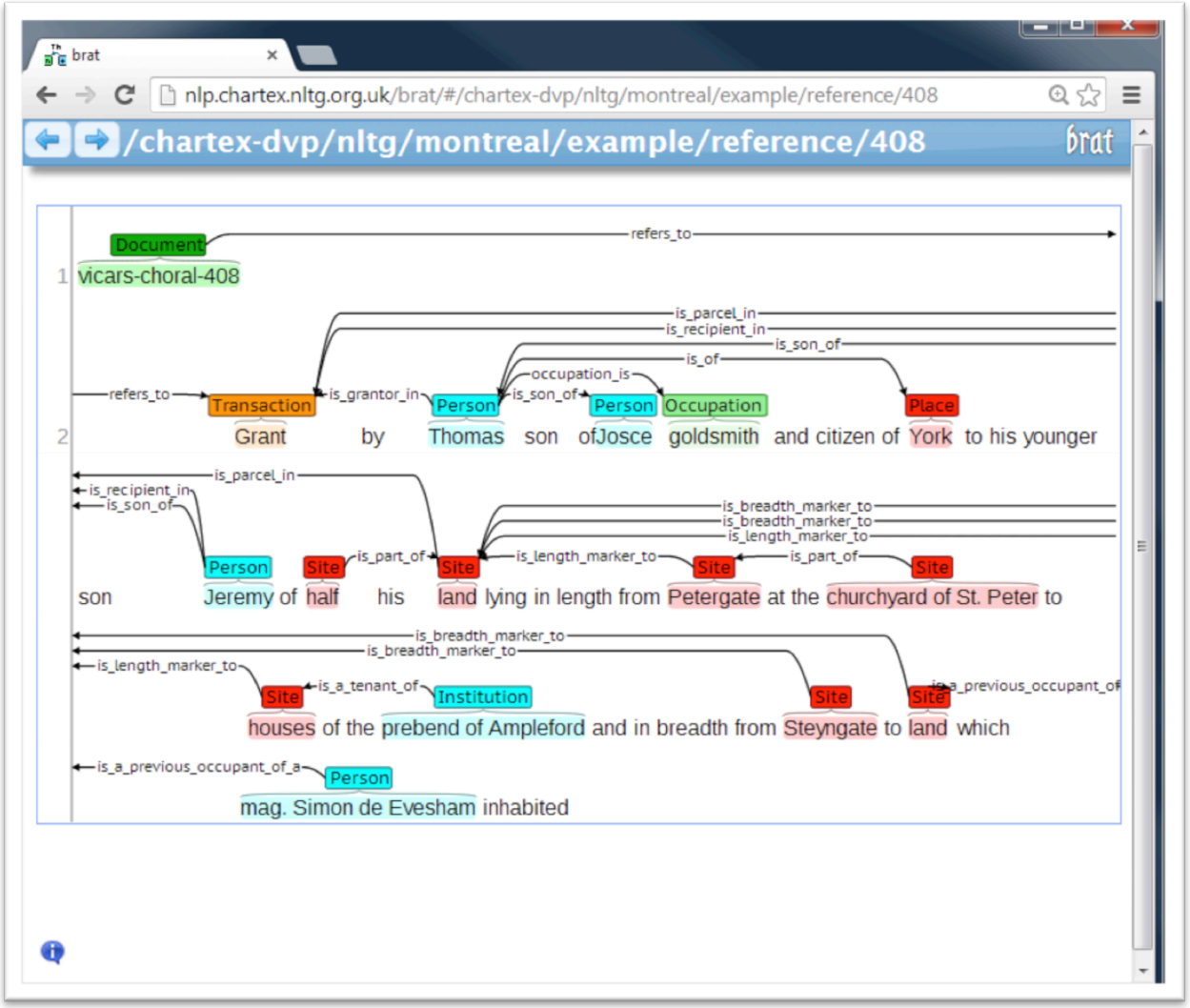
**Supporting domain experts to construct conceptual ontologies: A holistic approach**

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

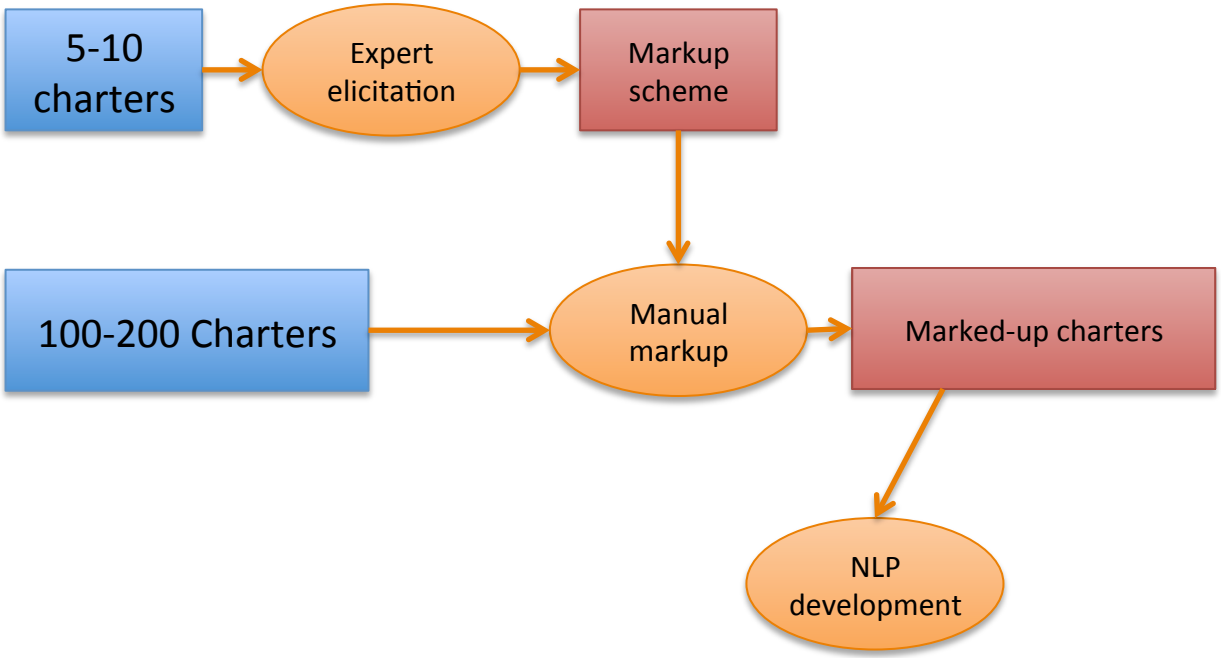


# BRAT rapid annotation tool

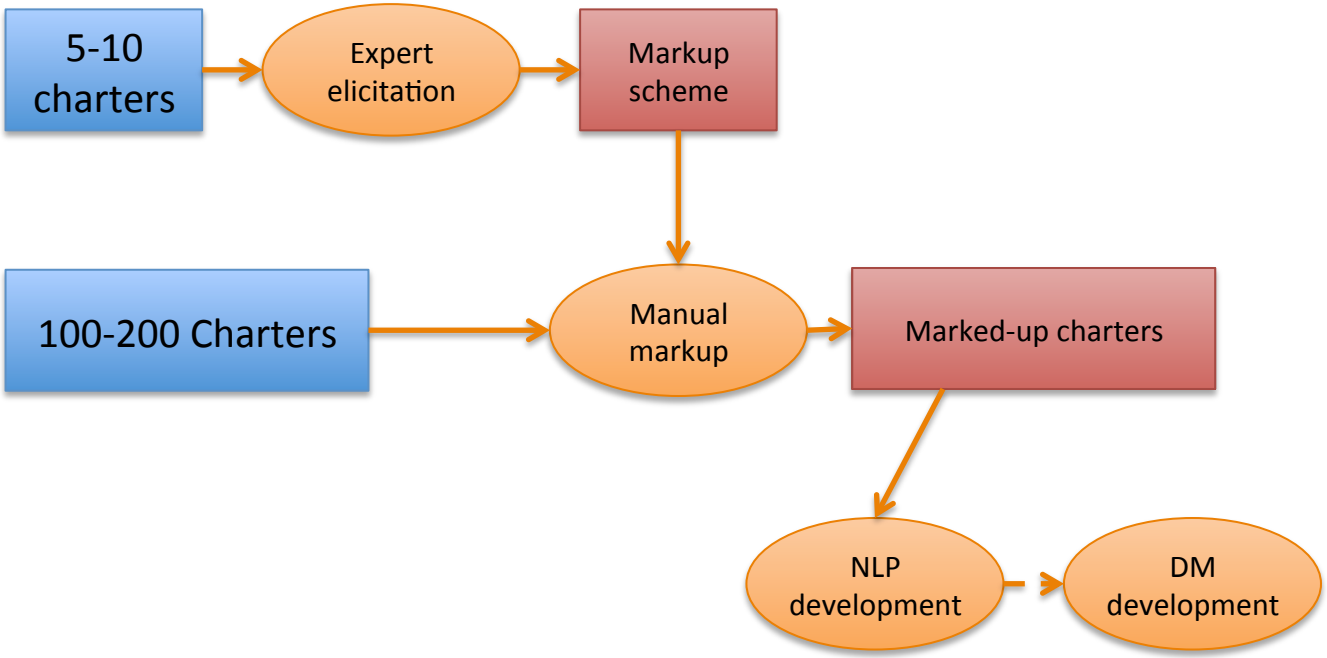



Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. (http://brat.nlplab.org/)

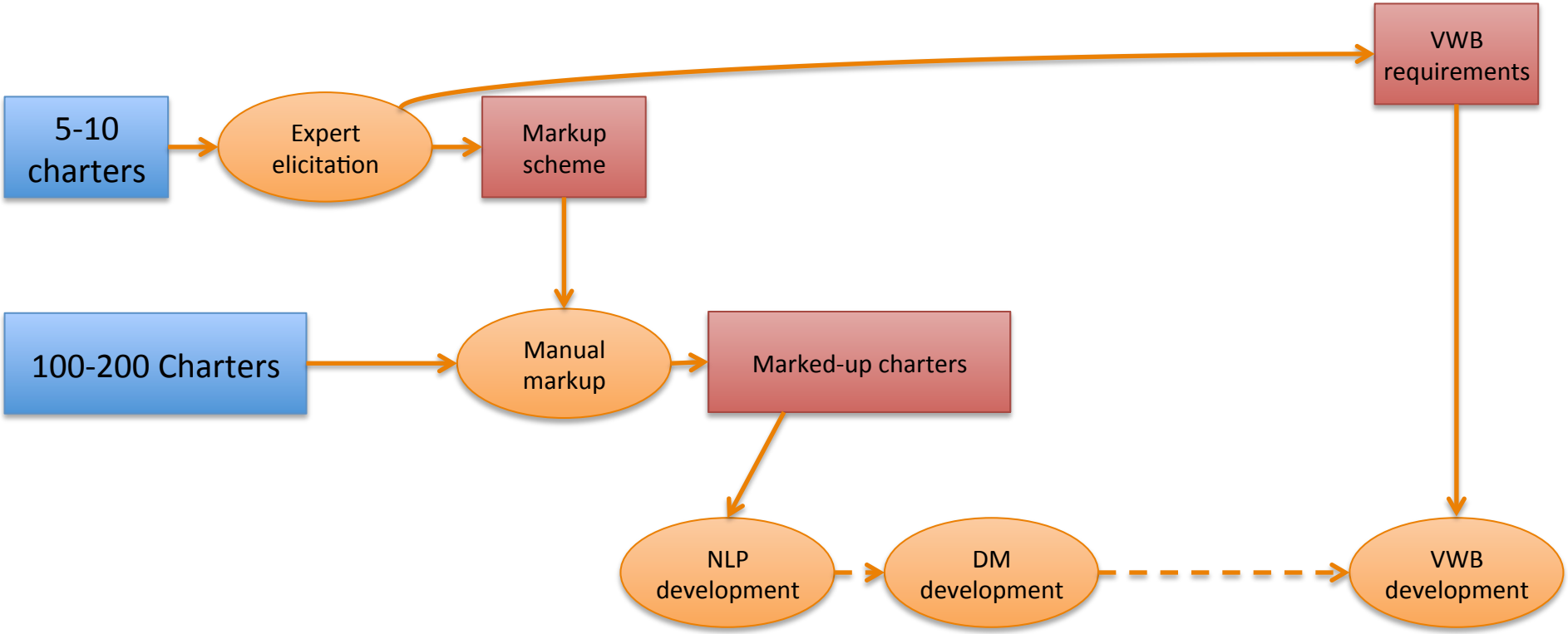
# ChartEx development process



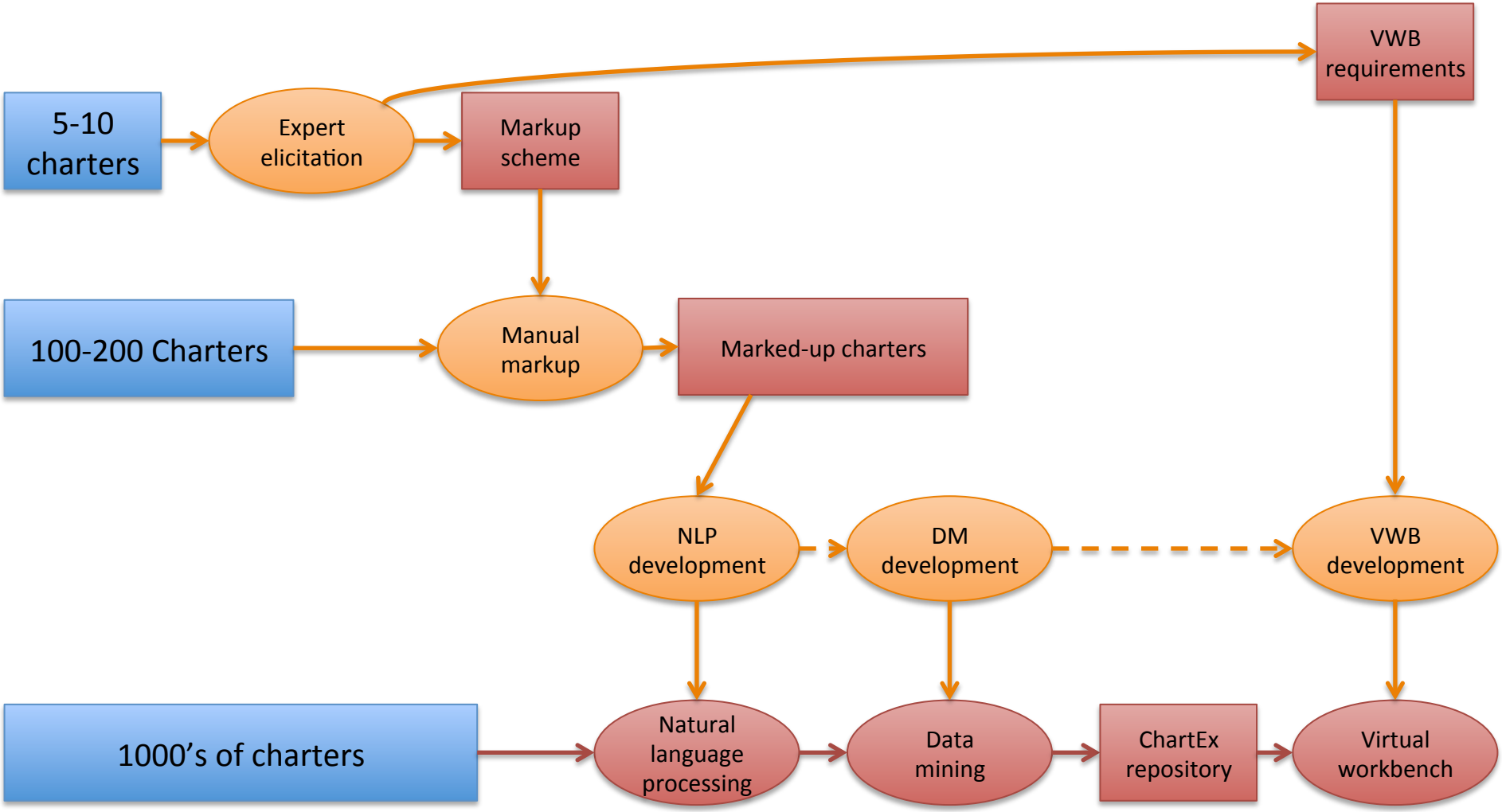
# ChartEx development process



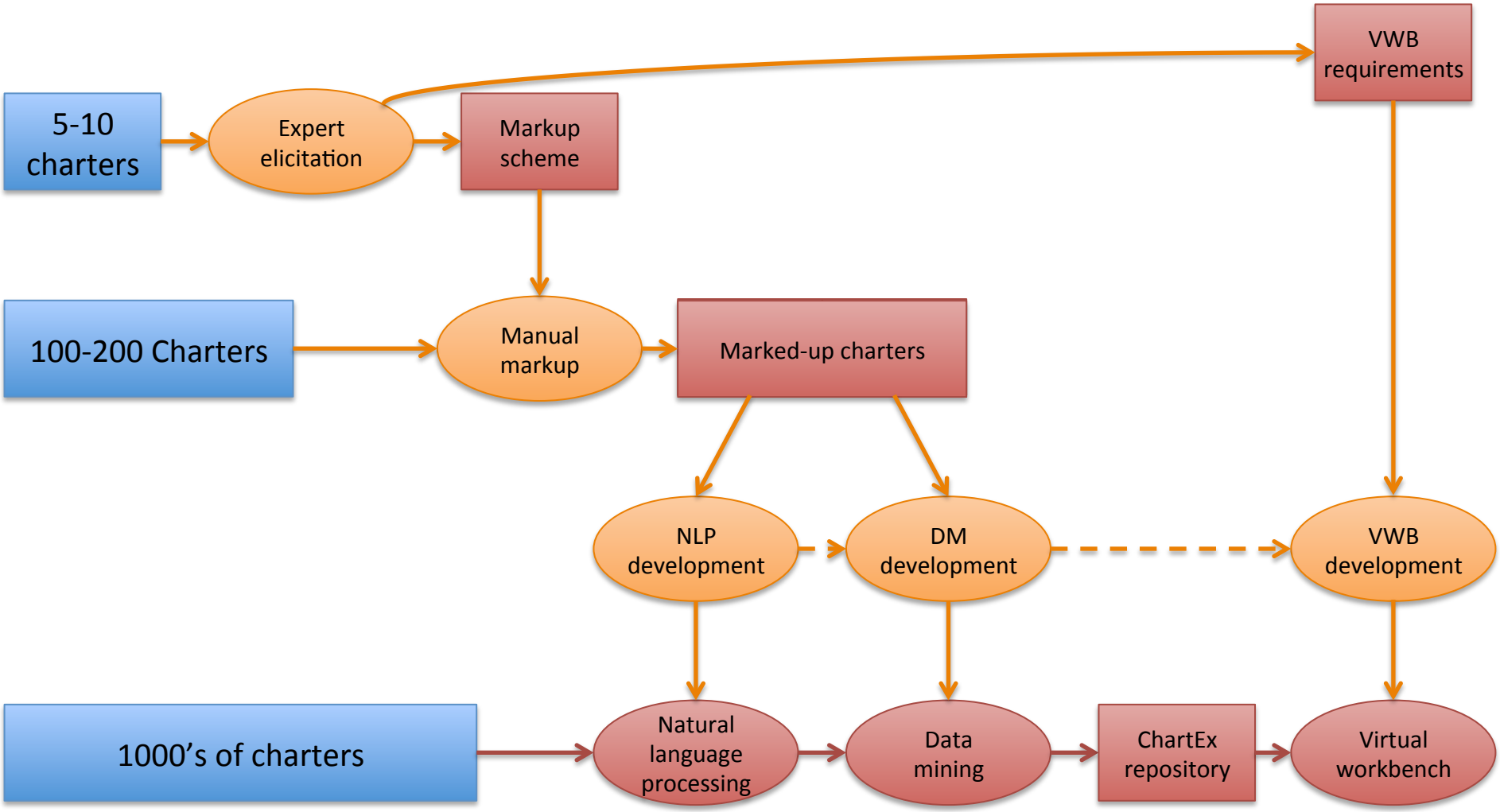
# ChartEx development process



# ChartEx development process



# ChartEx development process

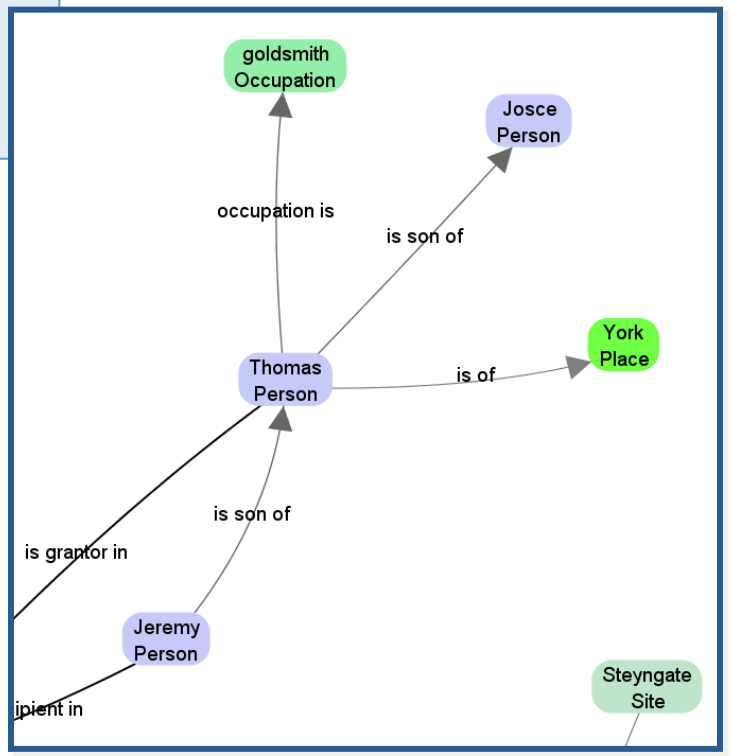


# The NLP task

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

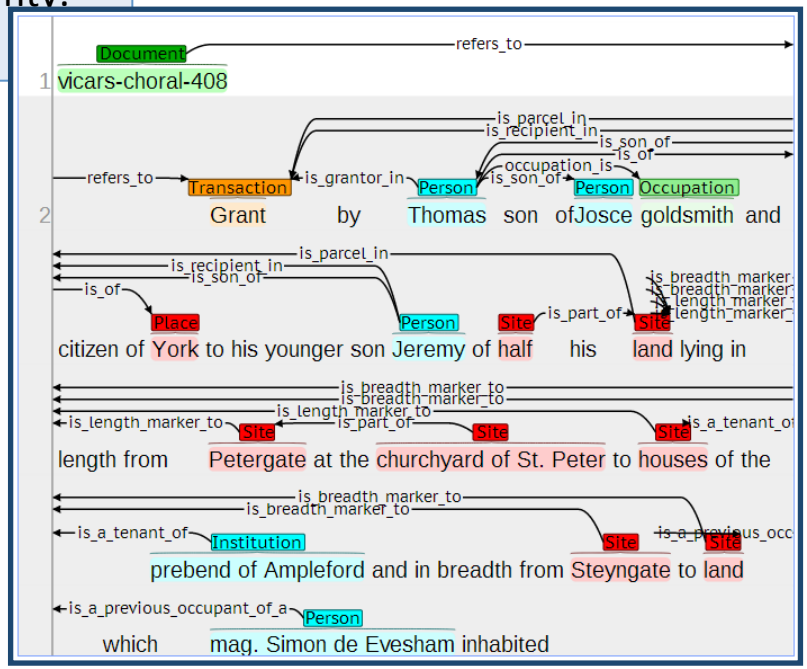


# The NLP task

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)



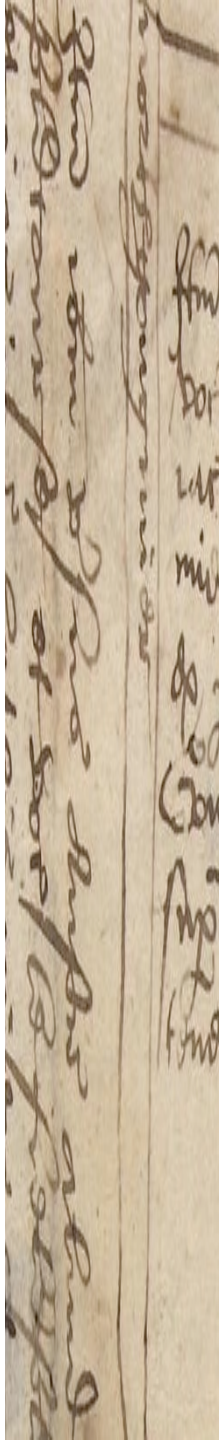
# The NLP task

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 …

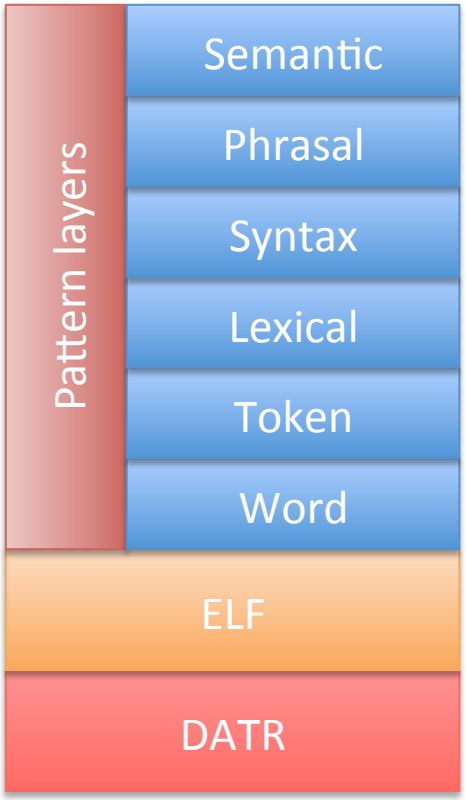
… although actually ‘under the hood’ it looks more like this

```
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.




Gluing phrases together using semantic expectations

Building phrases participating in subcat and coord etc

Building local syntactic structures

Identify multi-word expressions that behave like words

Identify properties of individual words (known types etc)

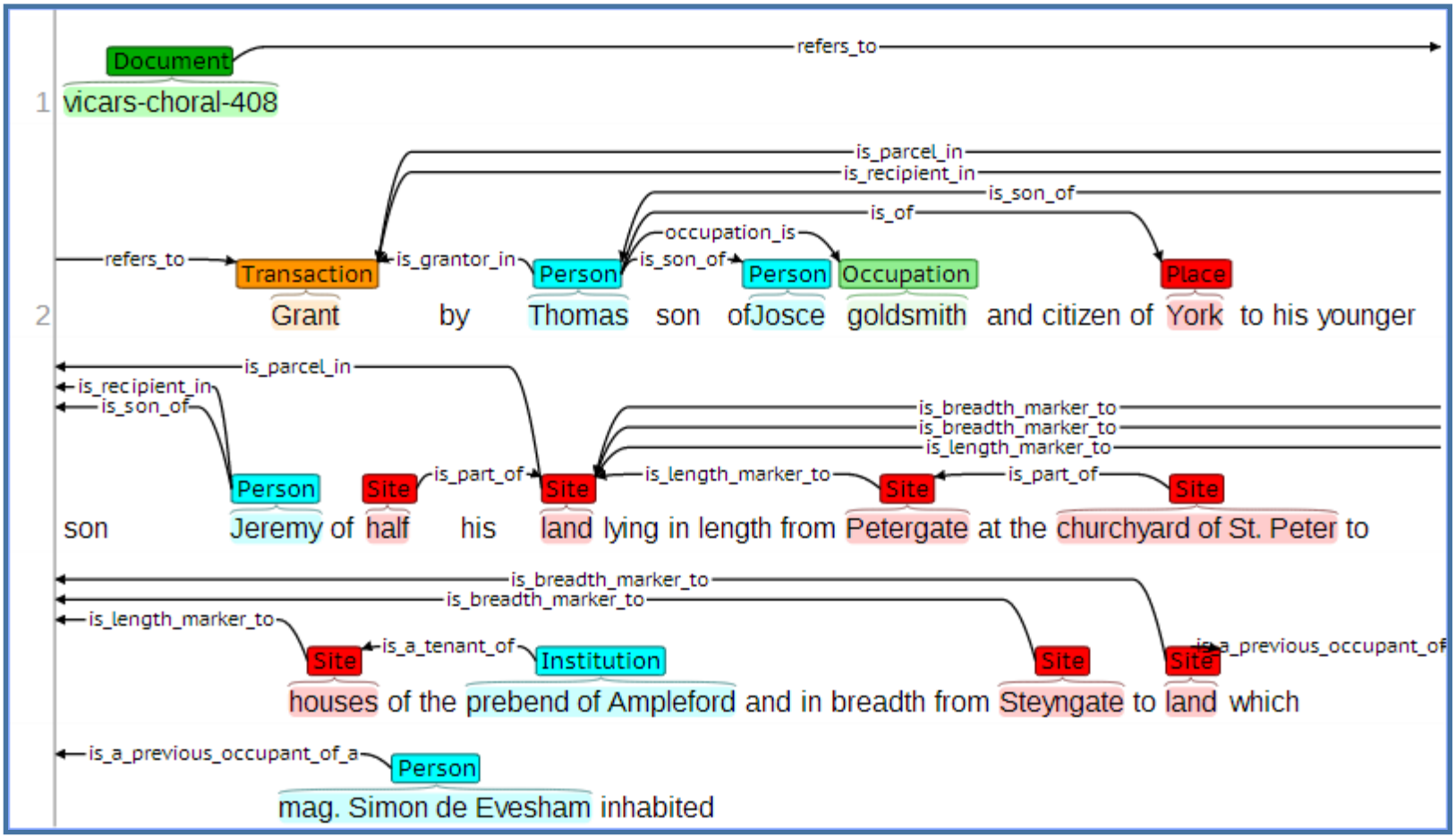
Bootstrapping lexical access using external tools (POS etc)

A lexical approach to multi-word/sentence processing

A default inheritance language for defining lexicons

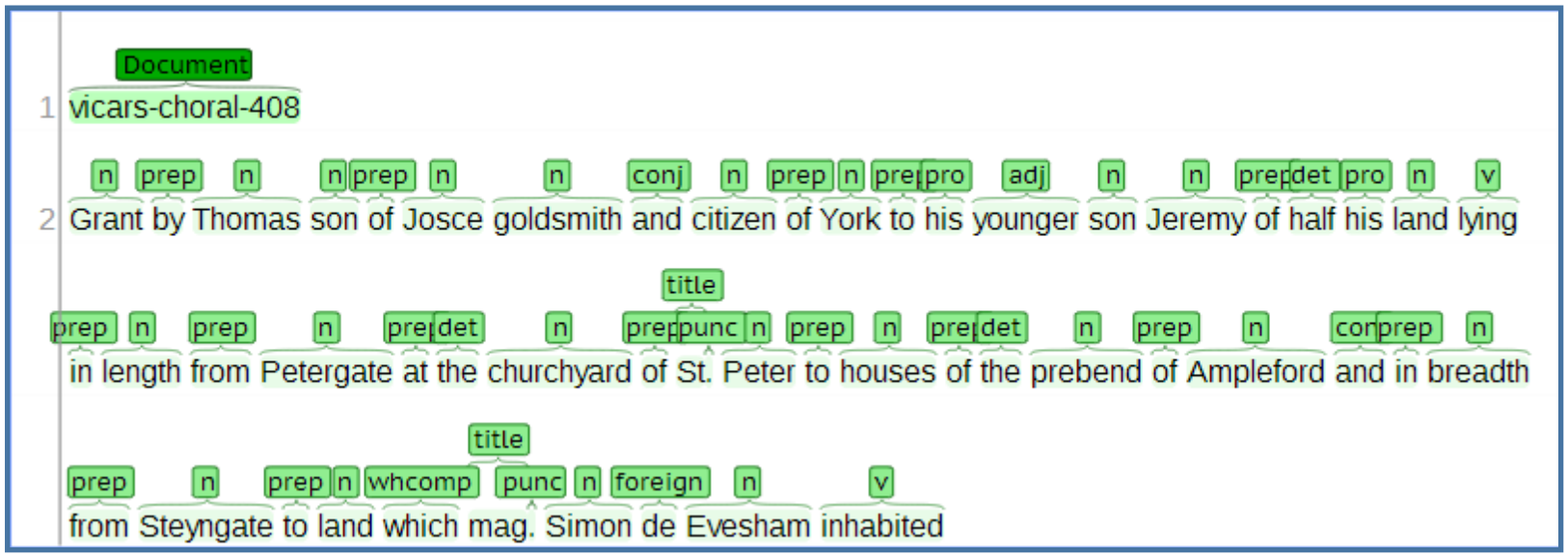
# The manual annotation

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;



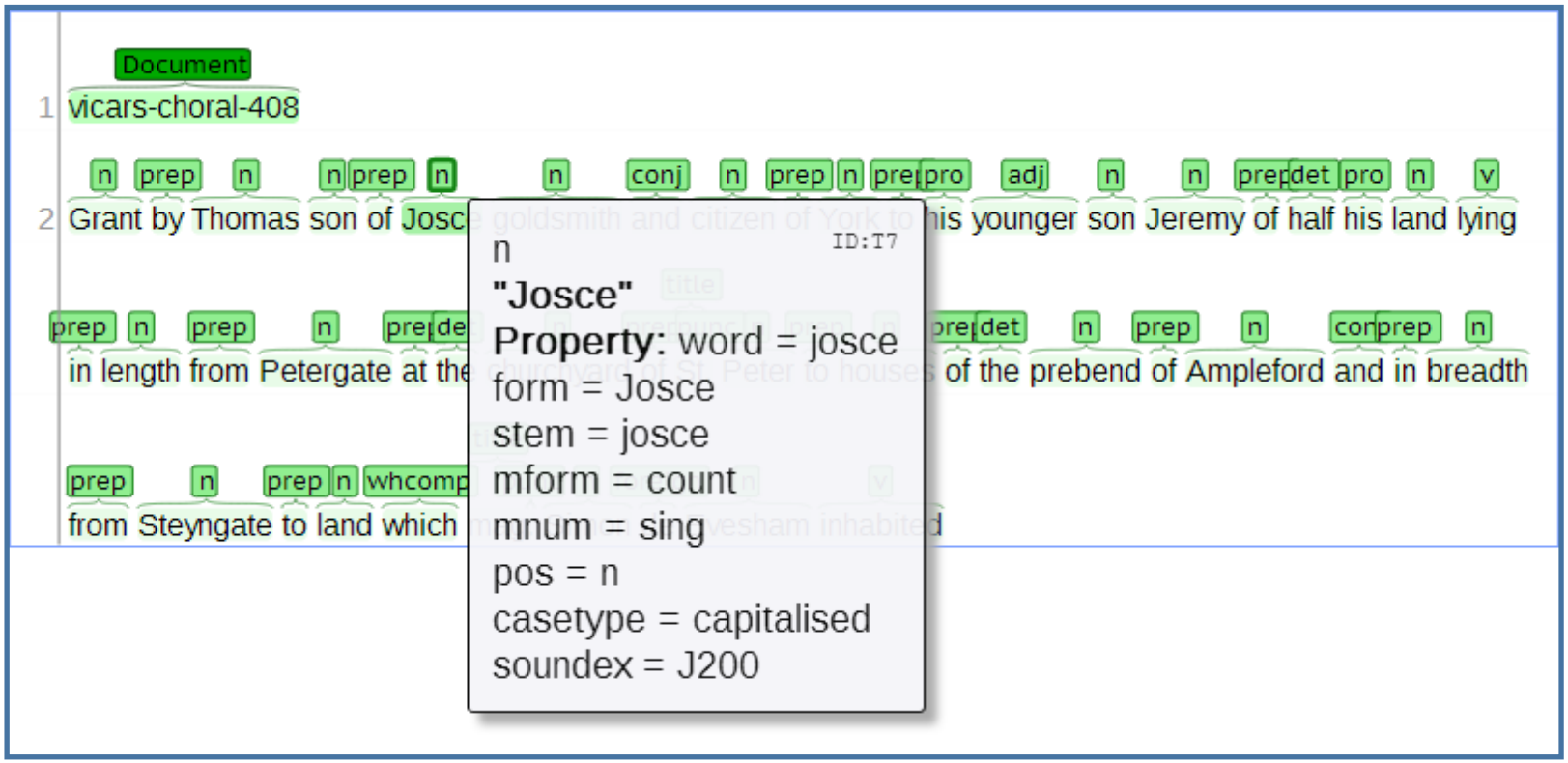
# NLP – layered pattern matching

**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.)



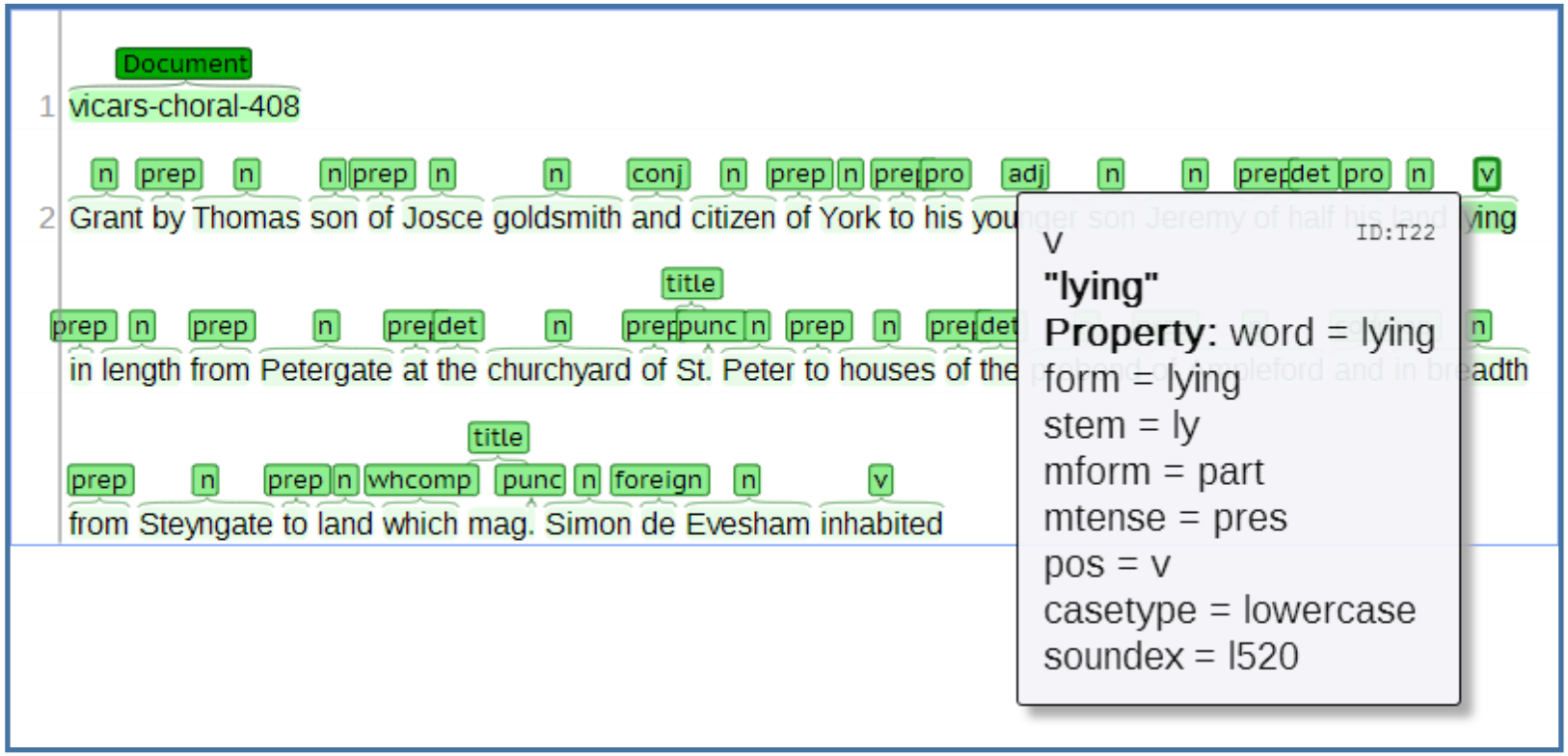
# NLP – layered pattern matching

**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.)



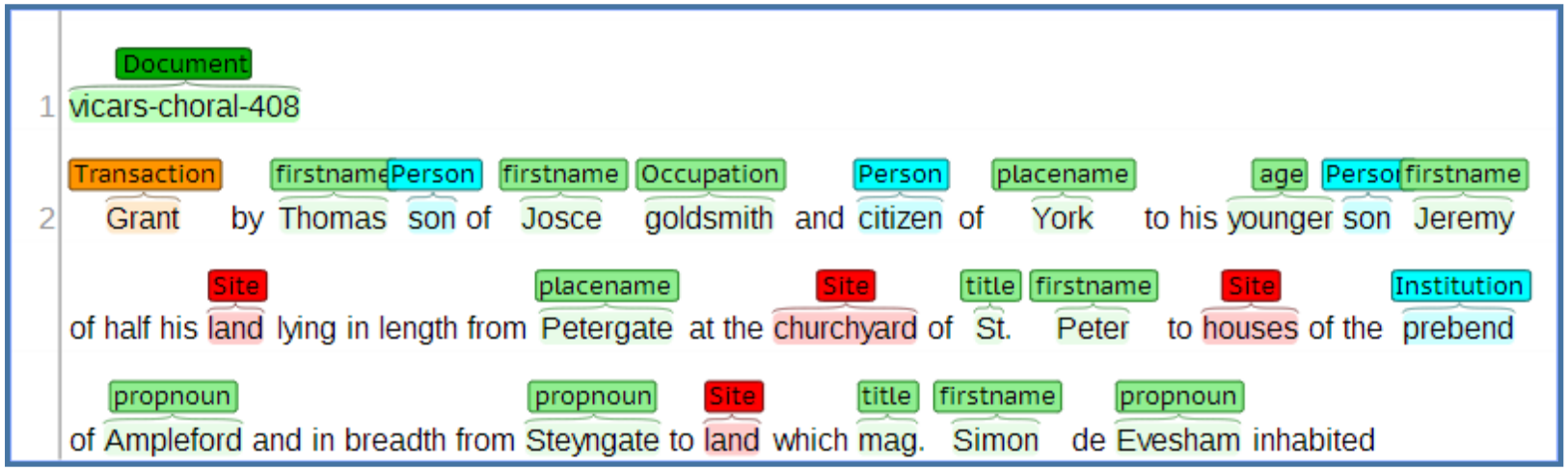
# NLP – layered pattern matching

**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.)



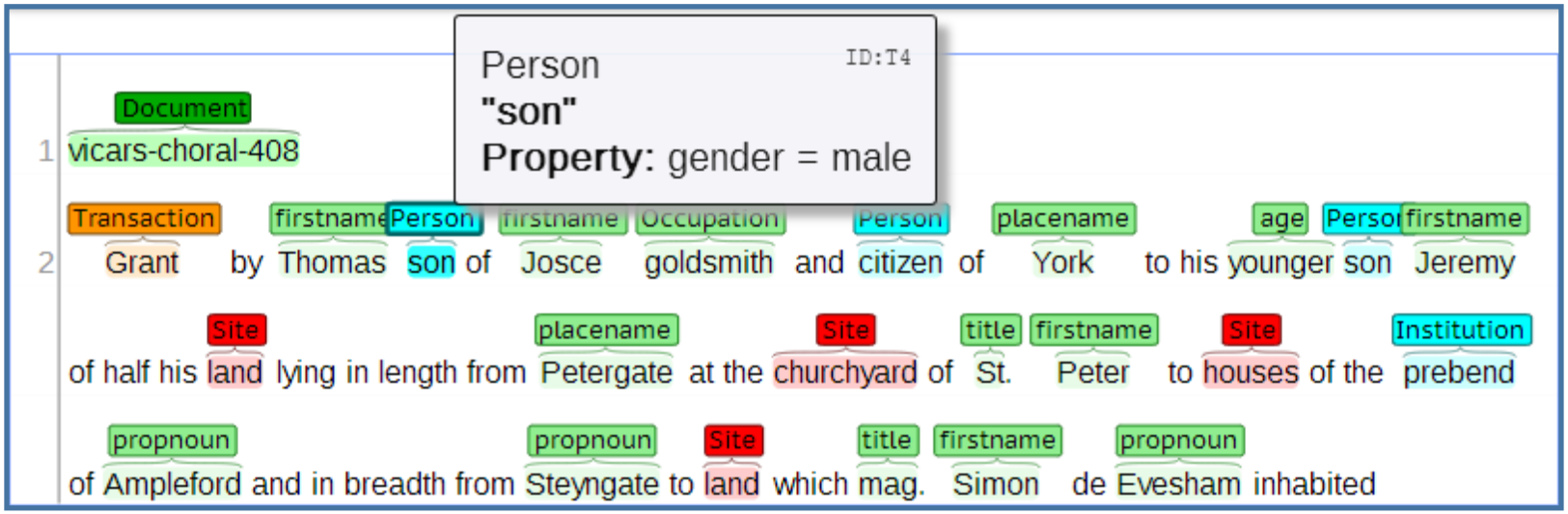
# NLP – layered pattern matching

**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.



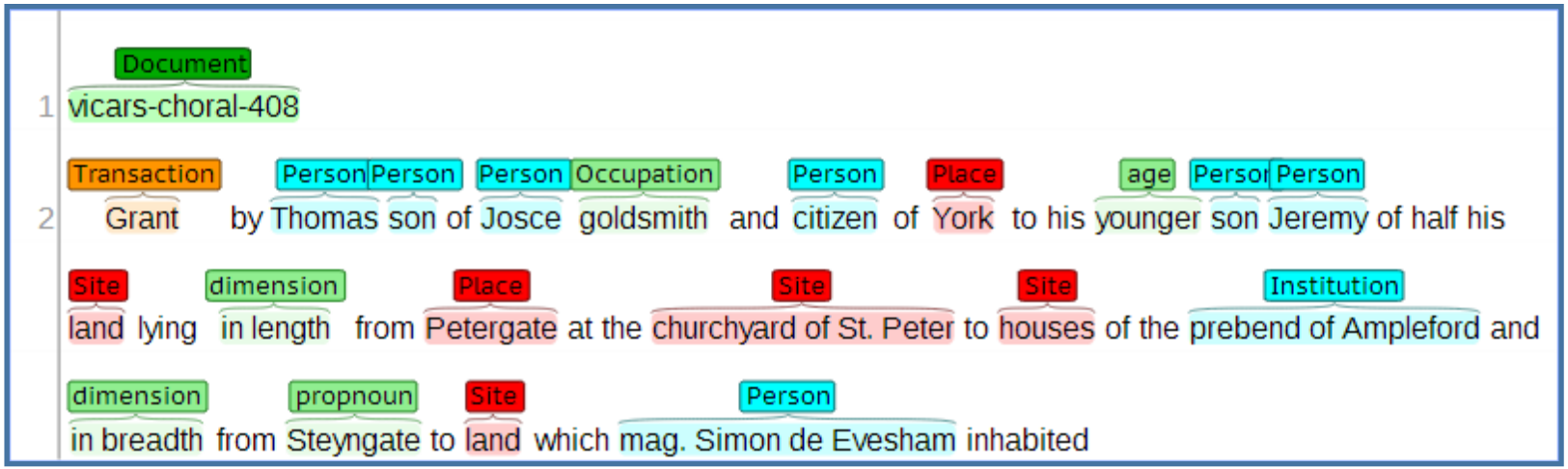
# NLP – layered pattern matching

**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.



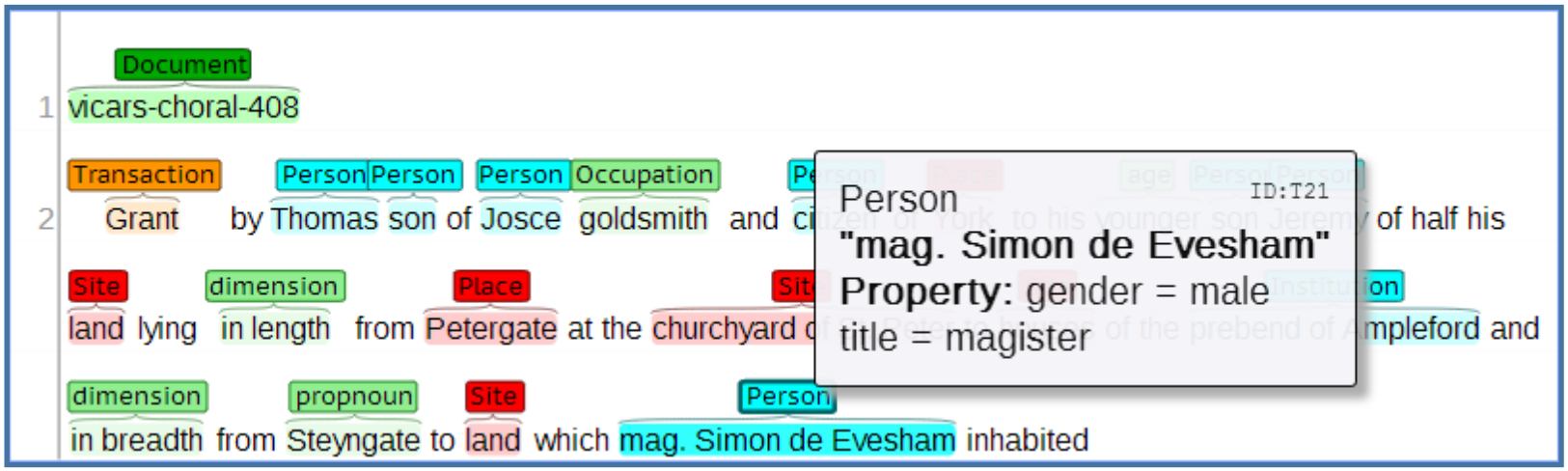
# NLP – layered pattern matching

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



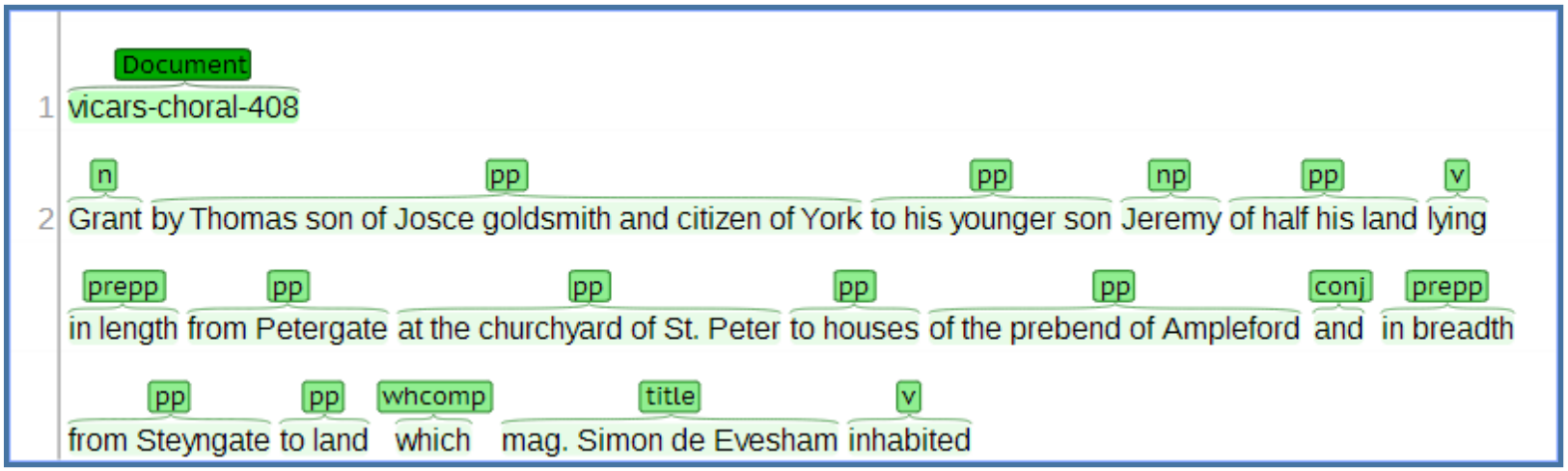
# NLP – layered pattern matching

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



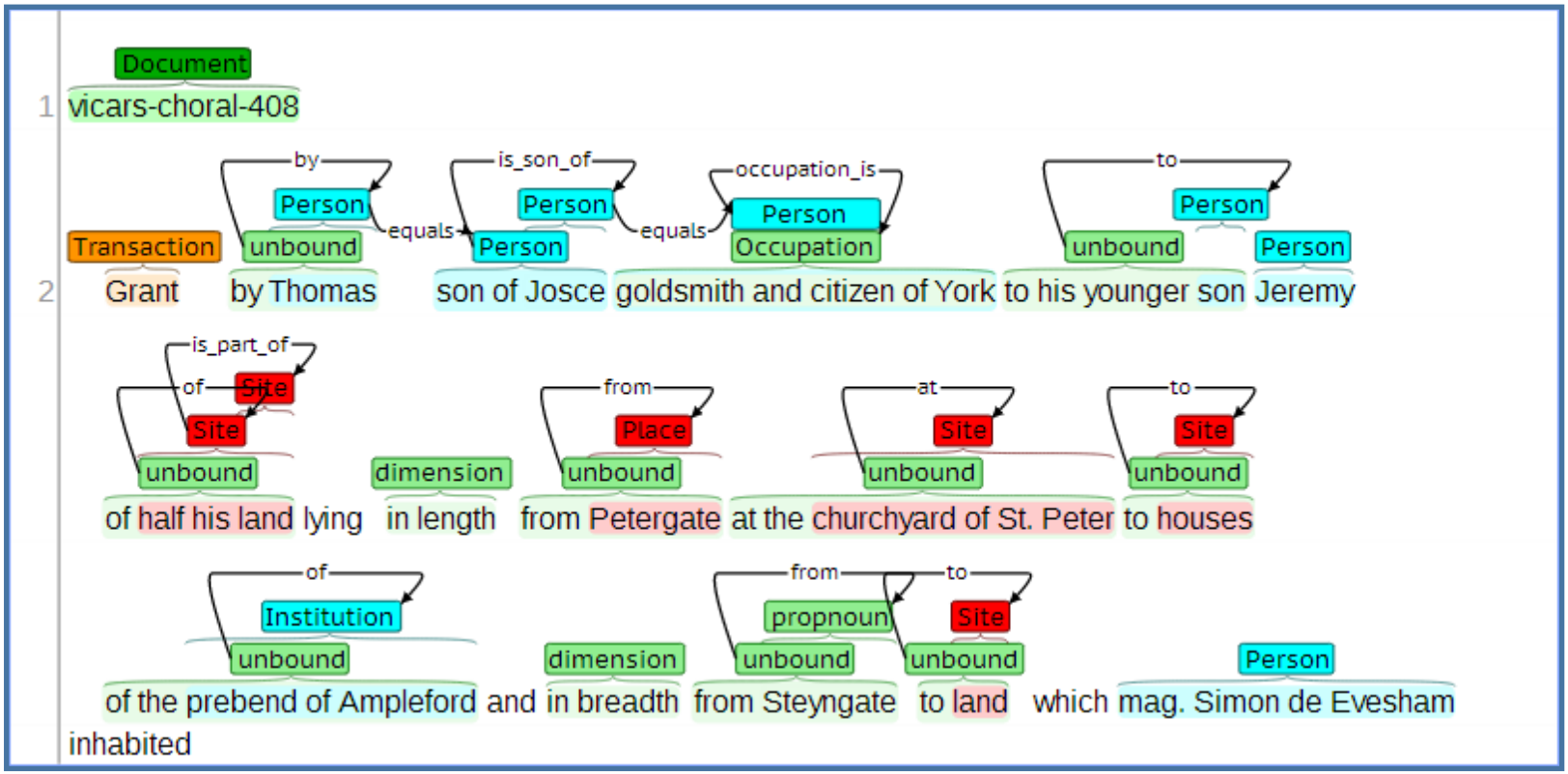
# NLP – layered pattern matching

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



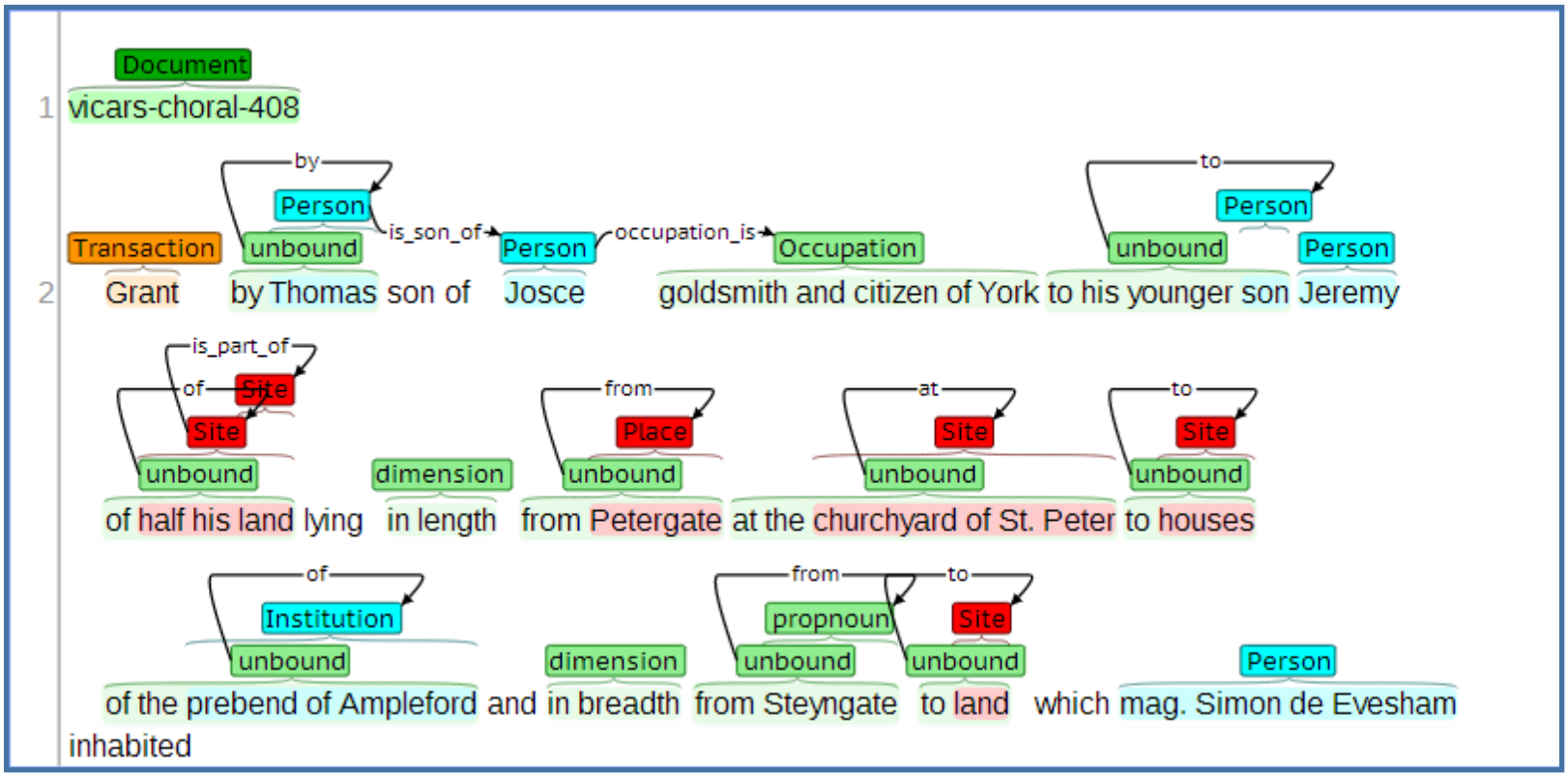
# NLP – layered pattern matching

**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.



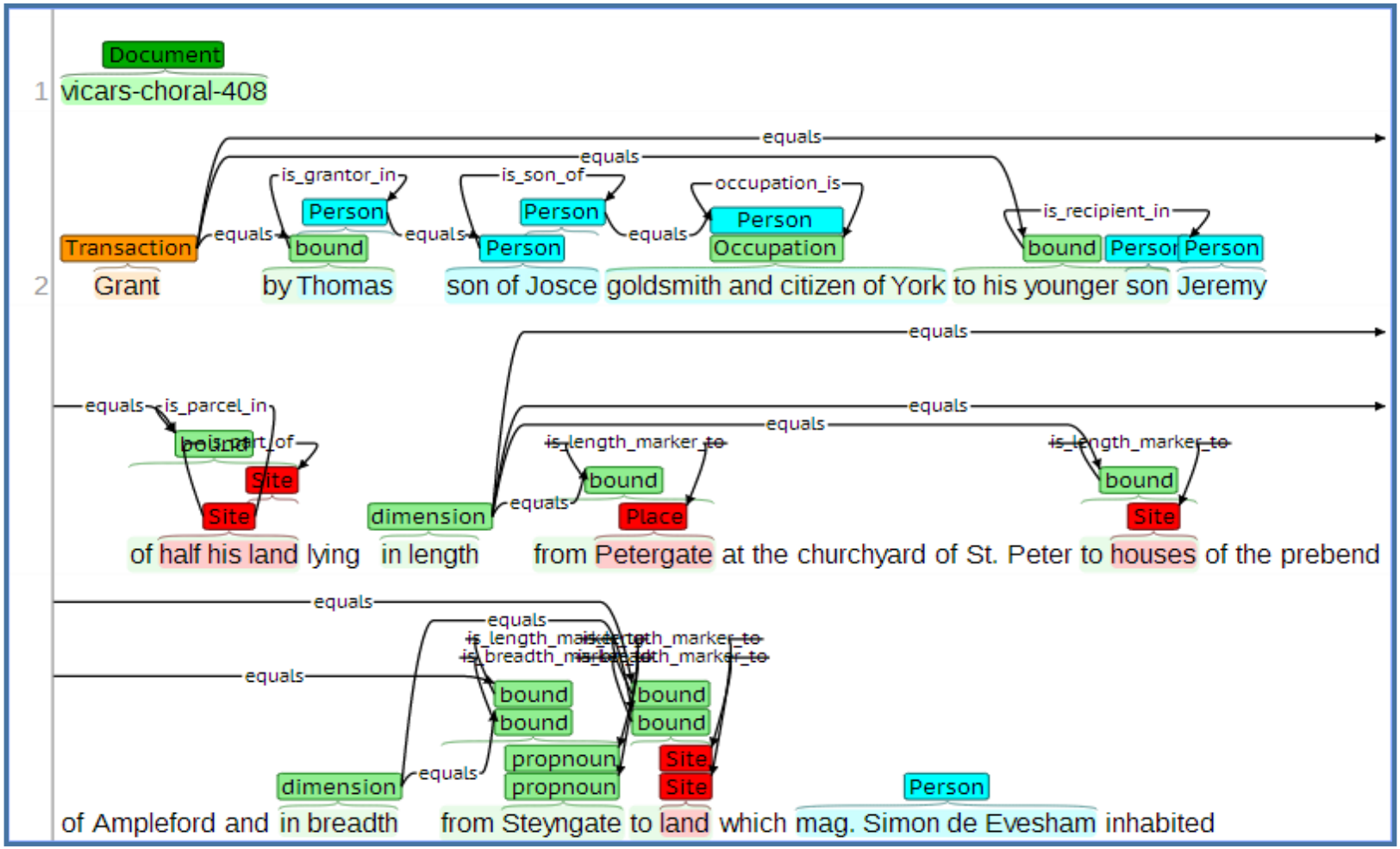
# NLP – layered pattern matching

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



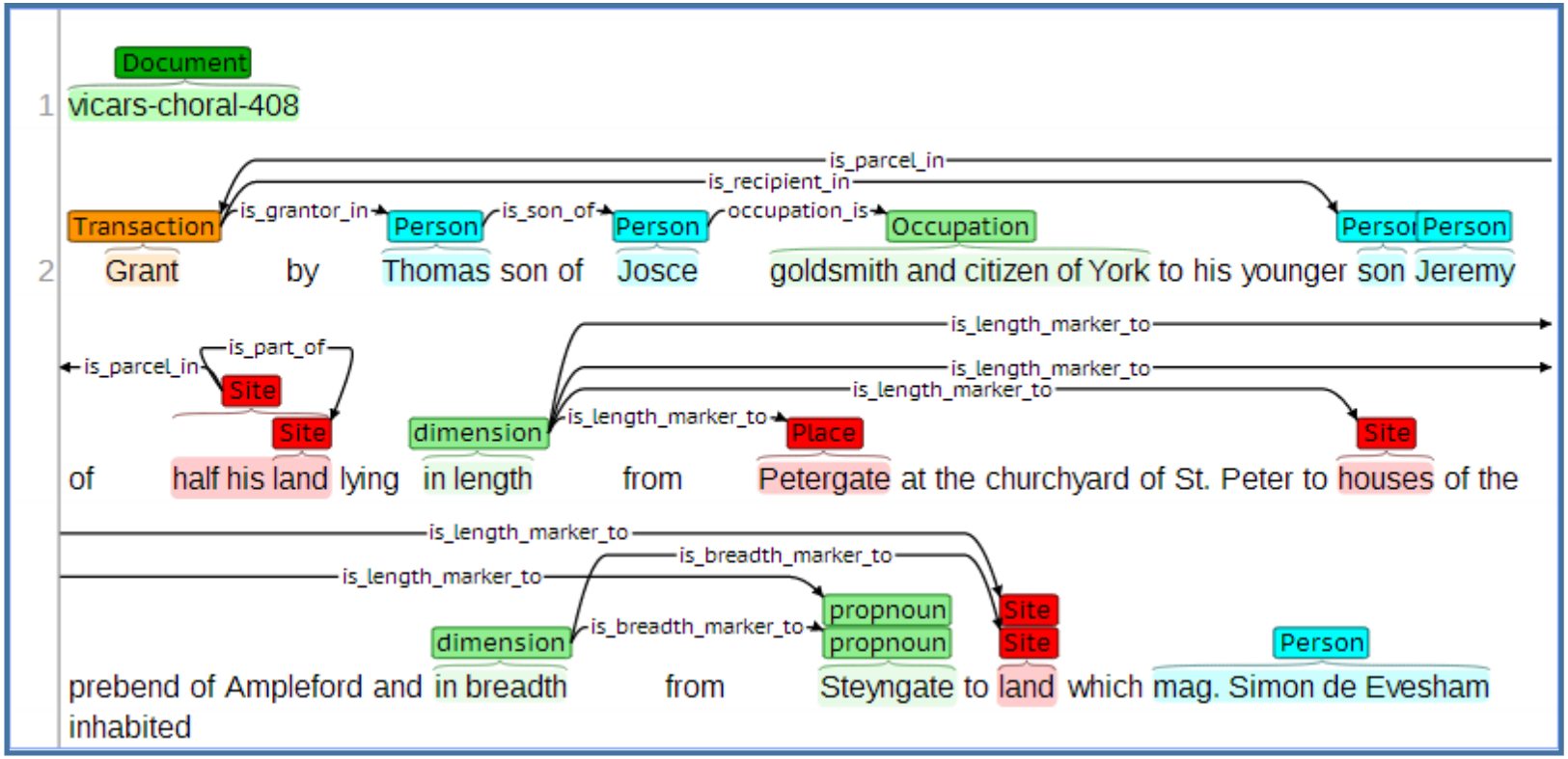
# NLP – layered pattern matching

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

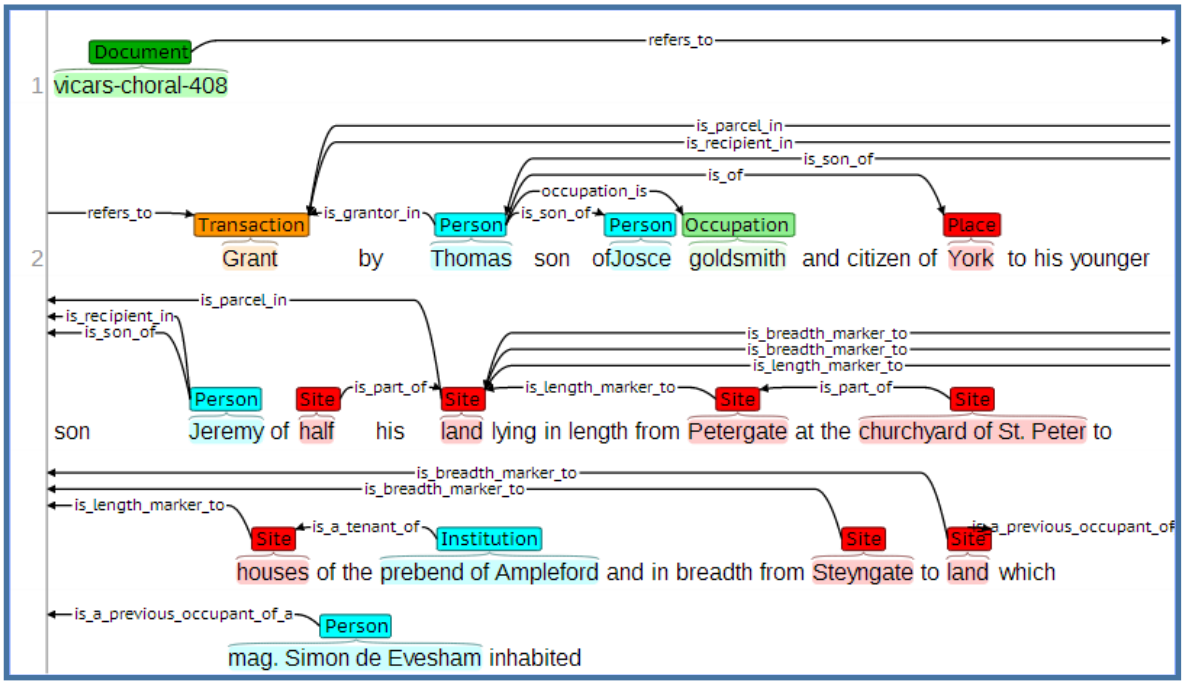


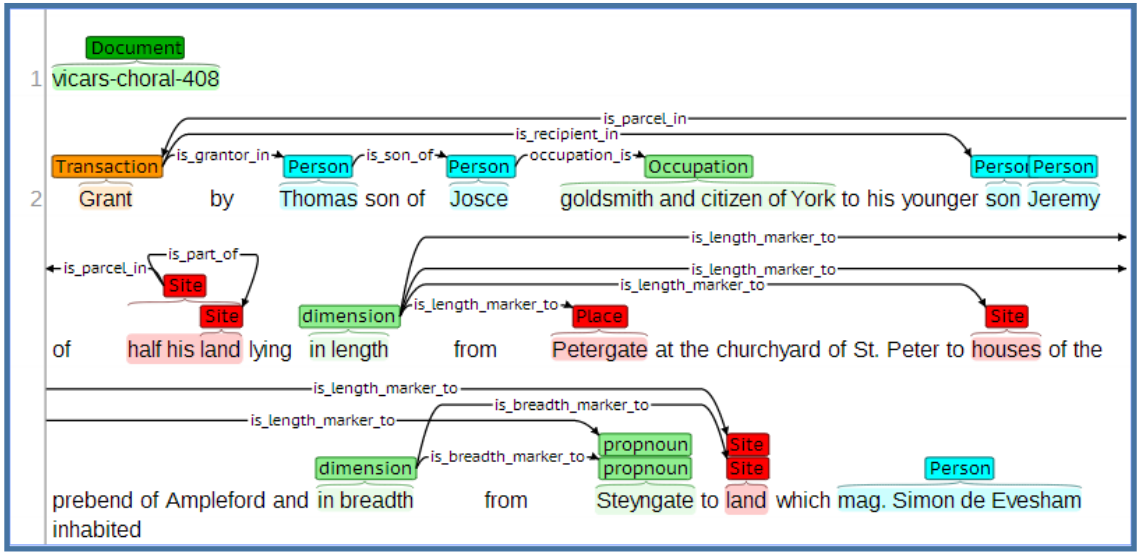
# NLP – layered pattern matching

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

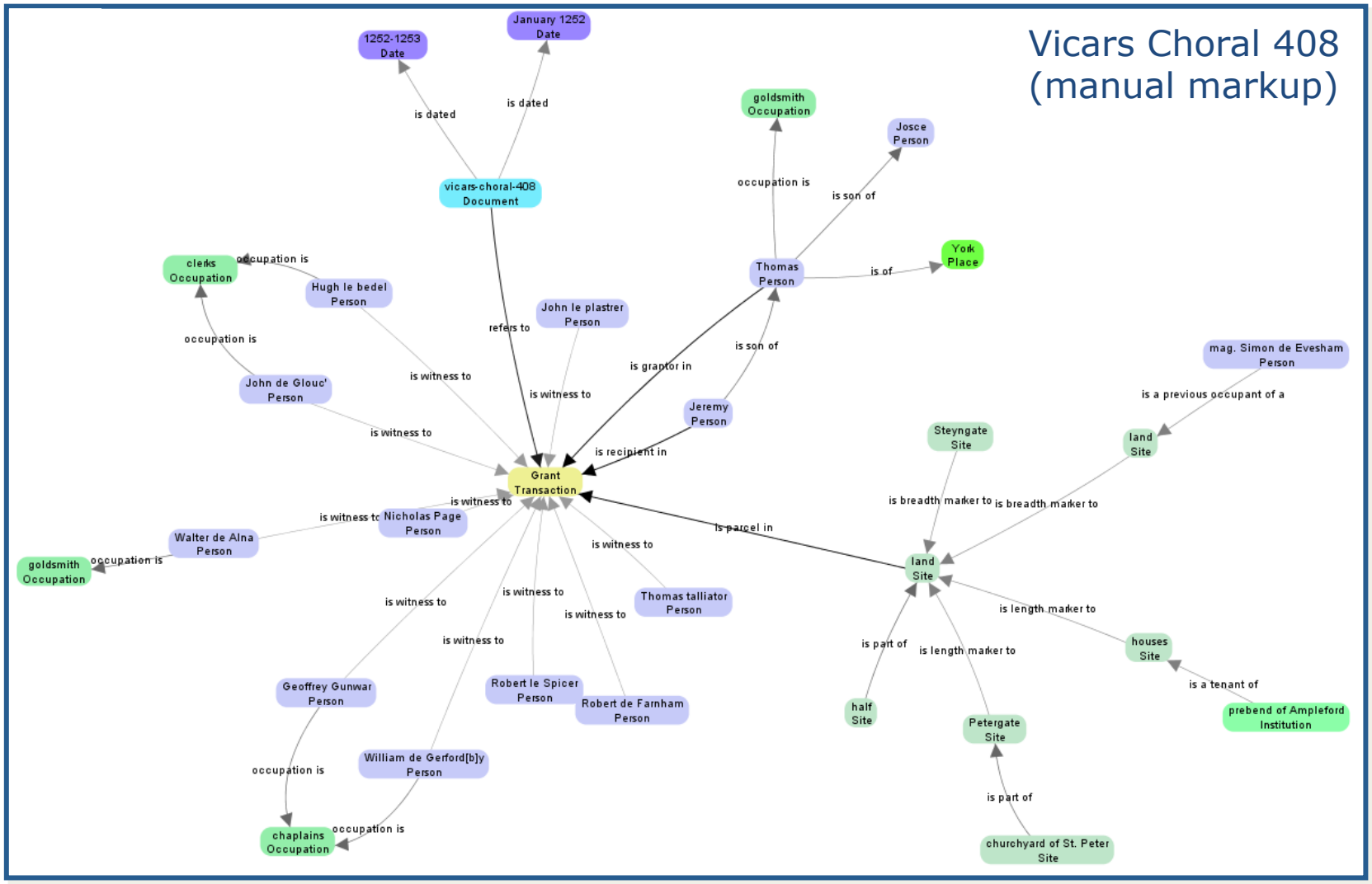


# How did we do?

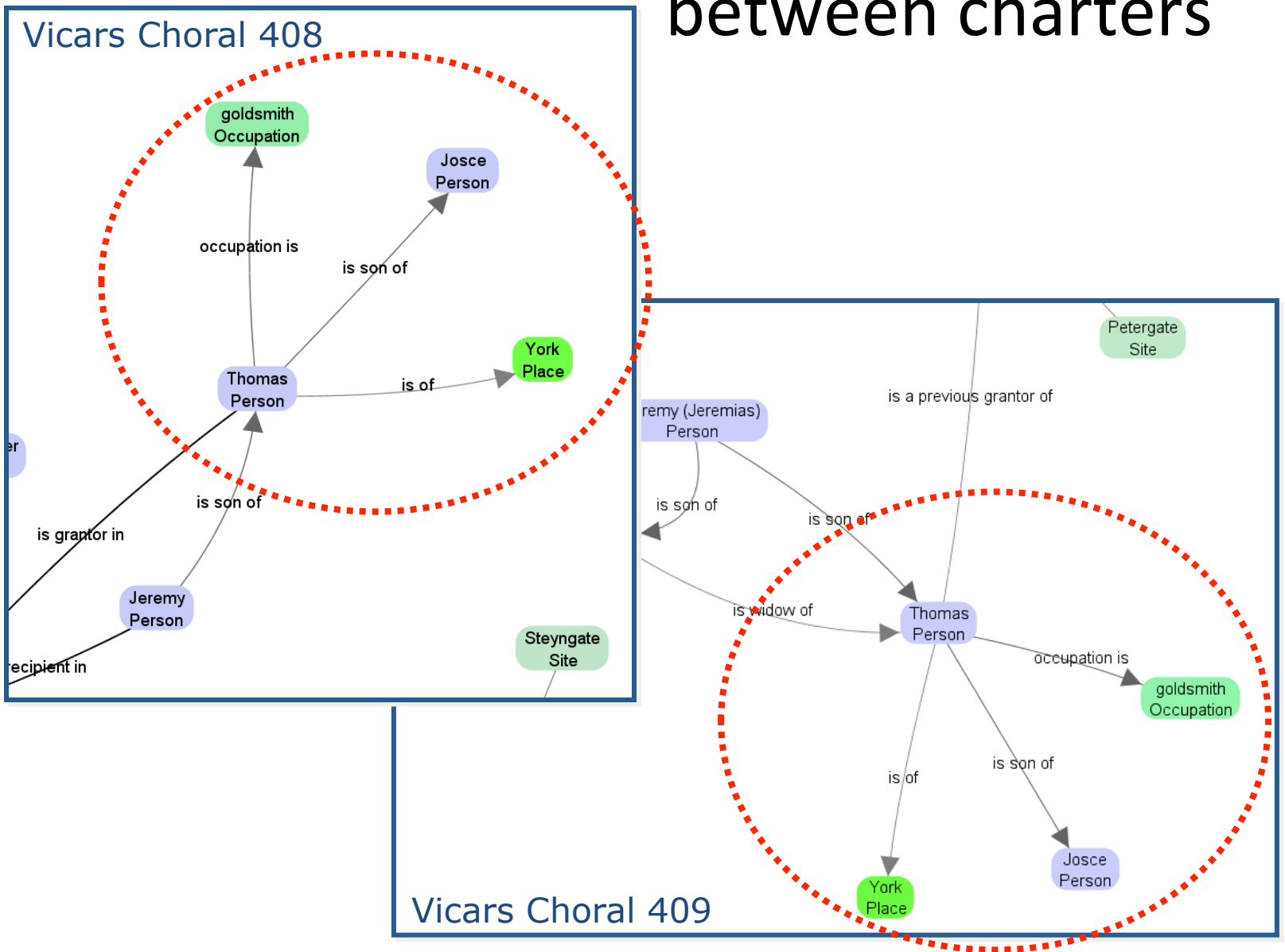

Document
refers_to
1 vicars-choral-408
is_parcel_in
is_recipient_in
is_son_of
is_of
occupation_is
refers_to
Transaction
is_grantor_in
Person
is_son_of
Person
Occupation
Place
2 Grant by Thomas son of Josce goldsmith and citizen of York to his younger
is_parcel_in
is_recipient_in
is_son_of
is_breadth_marker_to
is_breadth_marker_to
is_length_marker_to
Person
Site
is_part_of
Site
is_length_marker_to
Site
is_part_of
Site
son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to
is_breadth_marker_to
is_breadth_marker_to
is_length_marker_to
Site
is_a_tenant_of
Institution
Site
Site
is_a_previous_occupant_of
houses of the prebend of Ampleford and in breadth from Steyngate to land which
is_a_previous_occupant_of_a
Person
mag. Simon de Evesham inhabited


Document
1 vicars-choral-408
is_parcel_in
is_recipient_in
Transaction
is_grantor_in
Person
is_son_of
Person
occupation_is
Occupation
Person
Person
2 Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy
is_length_marker_to
is_parcel_in
is_part_of
Site
Site
is_length_marker_to
is_length_marker_to
dimension
is_length_marker_to
Place
Site
of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the
is_length_marker_to
is_length_marker_to
is_breadth_marker_to
propnoun
Site
dimension
is_breadth_marker_to
propnoun
Site
Person
prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham
inhabited

# Data Mining – probabilistic record linkage



# Matching information between charters

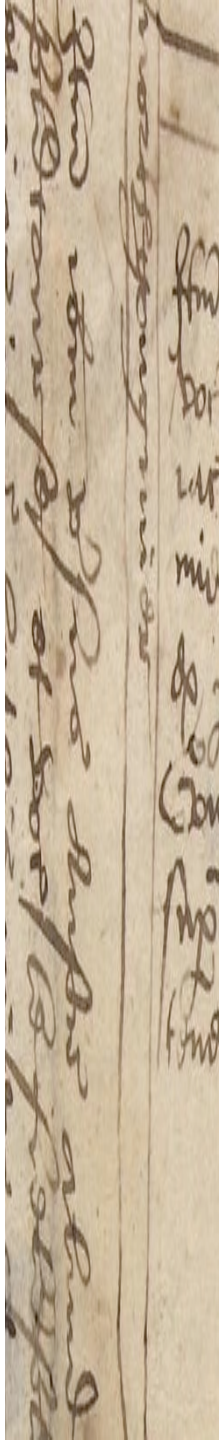


# Probabilistic reasoning
# “Thomas son of Josce, goldsmith”

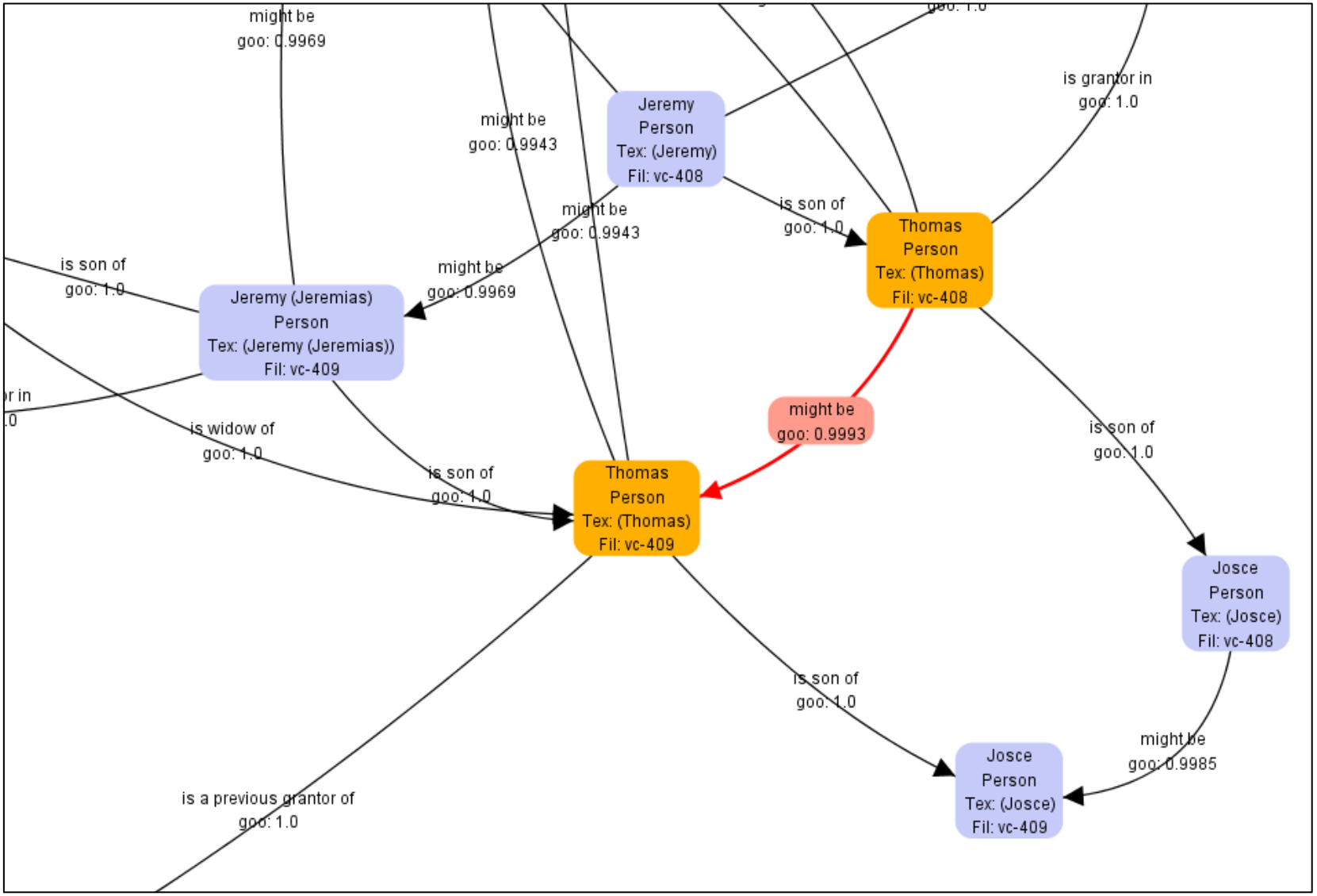
- Statistics
  *p*(Thomas) = 0.12 (common name)
  *p*(Josce) = 0.0015 (uncommon name)
  *p*(Goldsmith) = 0.04 (common profession)

- Dating
  vc-408 1252-1253
  vc-409 1253-1261

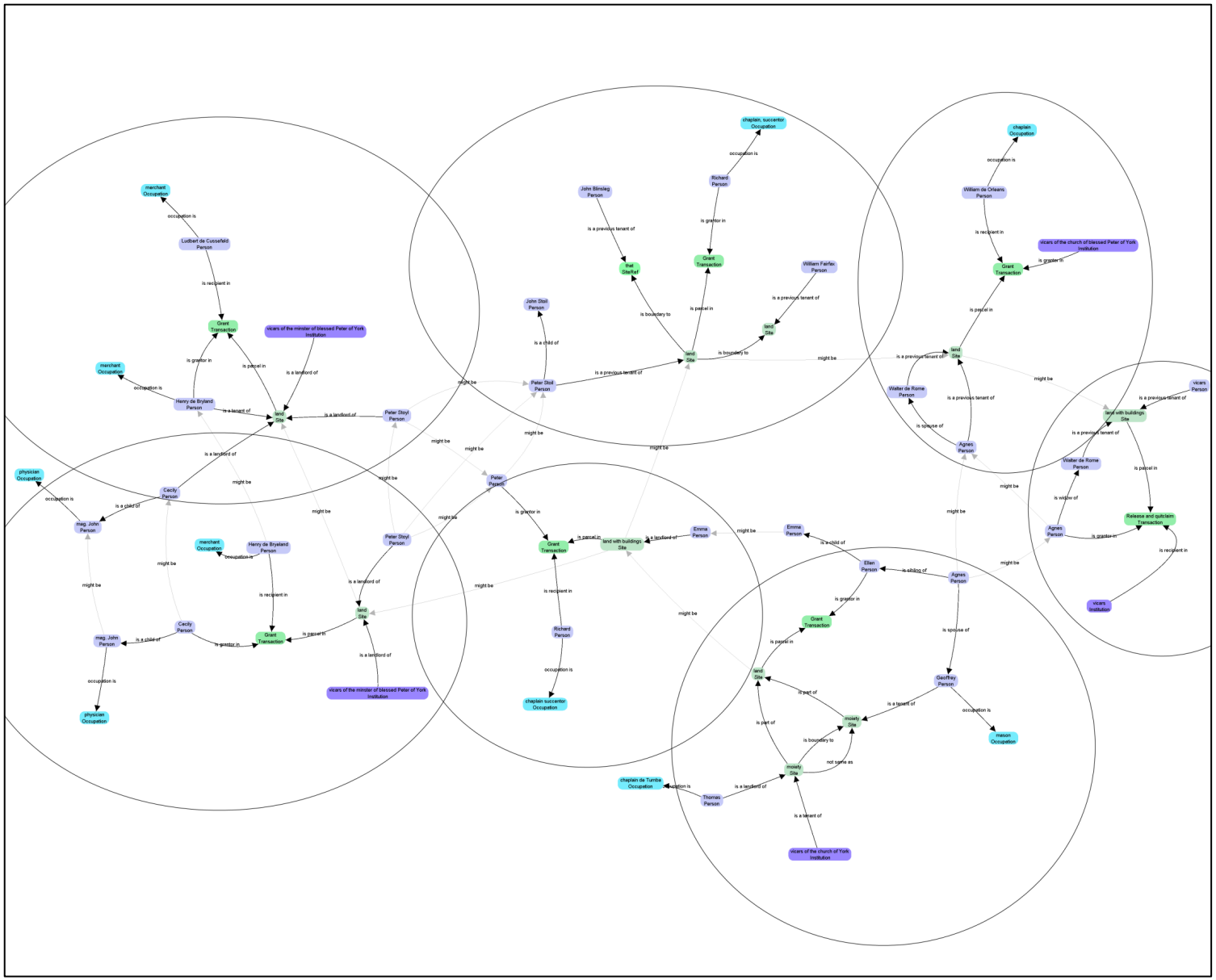
- Final confidence
  ***conf* (Thomas 408, Thomas 409) = 0.9993**



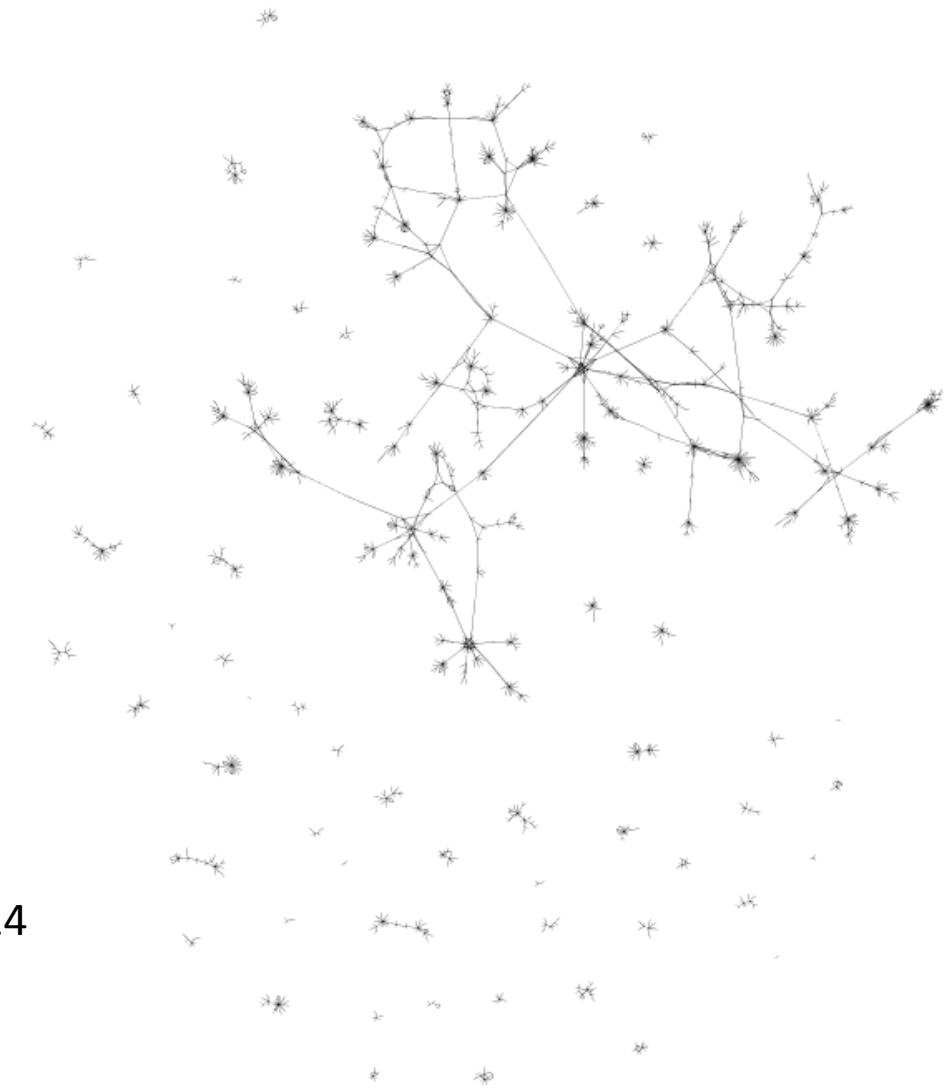
# “Thomas son of Josce” matched



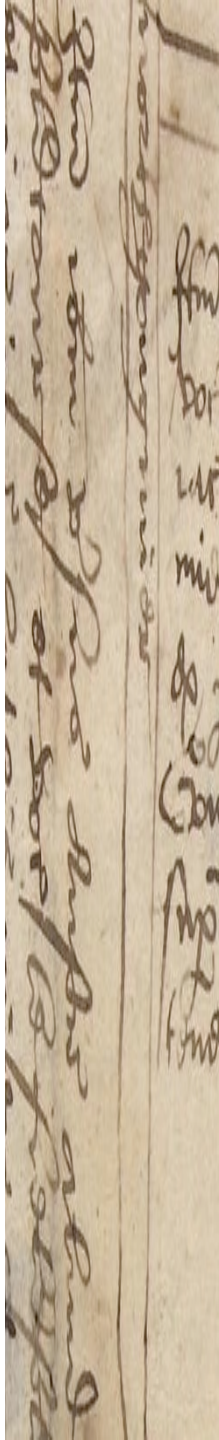
# Linking between charters



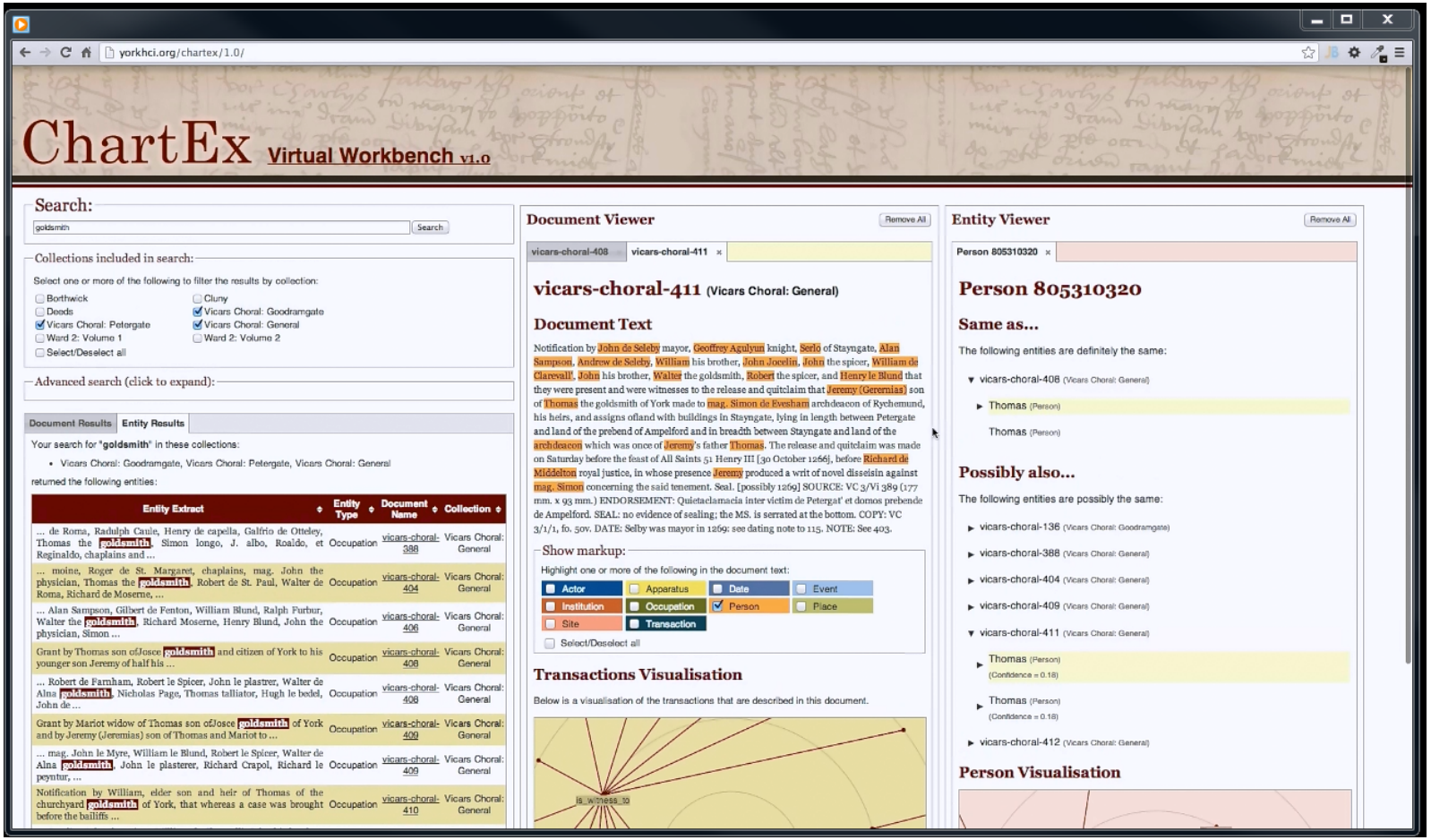
# Cartularies as networks



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

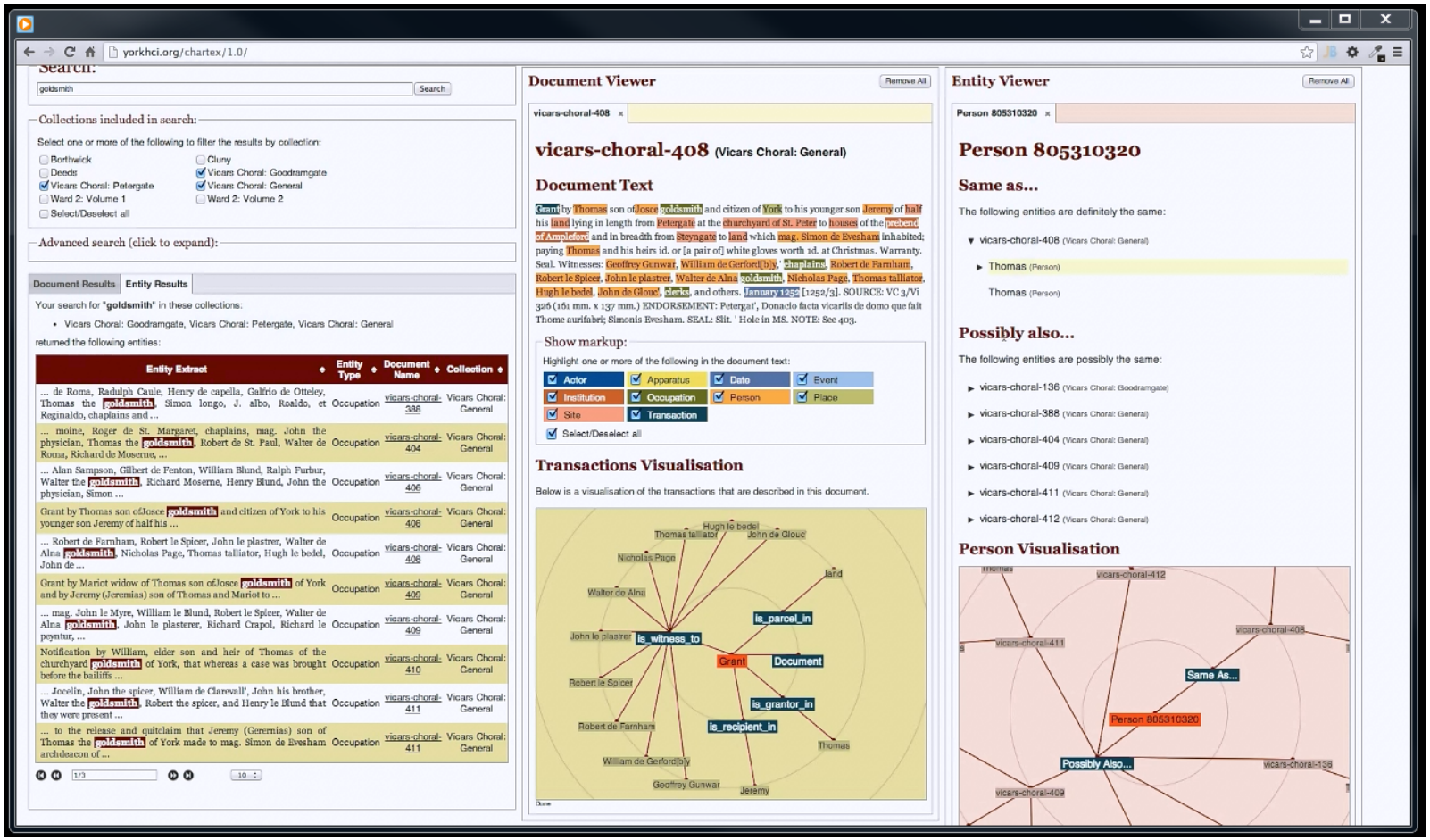


# The ChartEx Virtual Workbench




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

# The ChartEx Virtual Workbench




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

# 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.)