ChartEx: Discovering spatial descriptions and relationships in medieval charters

Introduction

www.chartex.org

@ChartExProject
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

www.diggingintodata.org
ChartEx Project Partners

**University of York:** History and 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 U of Toronto, Columbia Digital Humanities

**Advisory Panel:** various cognate projects
ChartEx Team

**History:** Sarah Rees Jones, Adam Kosto, Michael Gervers, Robin Sutherland-Harris, Stefania Perring, Jon Crump

**NLP:** Roger Evans, Lynne Cahill

**Data Mining:** Arno Knobbe, Marvin Meeng

**HCI:** Helen Petrie, Christopher Power, David Swallow
Digital Charter Collections


The National Archives (UK) Ward 2. Abridged English translations

The Borthwick Institute, Yarburgh Muniments. Abridged English translations

DEEDS, University of Toronto. Latin charters of English provenance.

Cluny. Latin charters of French provenance (CBMA)
Urban Historic Topography

York
corner of Stonegate and Petergate
opposite Minster Gates
detail of OS 1852
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[by], 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
Chronological Sequencing and Spatial Sequencing

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

409. Grant by Mariot widow of Thomas son of Josce goldsmith of York and by Jeremy (Jeremias) son of Thomas and Mariot to mag. Simon de Evesham canon of York of land with buildings in Steyngate, granted to them by Thomas, lying in length between Petergate and land of the prebend of Ampleford, and in breadth between Steyngate and land once of Geoffrey de Norwyz precentor of York
Conceptual plan of logical relationships

LAND and BUILDINGS
(prebind of Ampleford)

LAND WITH BUILDINGS
(Simon de Evesham
archdeacon, previously of
late Thomas son of
Josce; half land was
given to Jeremy, the
shop and solar on the
corner was of Mariot and
Jeremy)

PROPERTY 3

LAND AND HOUSES
(Simon de Evesham,
precentor,
previously of Geoffre
de Norwich precentor)

PROPERTY 2

LAND of the
prebind of
Barney

half

LAND AND
HOUSES (Simon
de Evesham
precentor
previously of
Anketin de
Hesingwald
and
Thomas Sotewain

PROPERTY 1

LAND and
BUILDINGS
(prebind of
Masham)

St. Peter’s
Churchyard

VCC 403-411 Property 1, 2, 3 and their development by 1266
ChartEx Architecture

- Charter documents
- Natural language processing
- Analysed individual documents
- Data mining
- Analysed integrated documents
- ChartEx workbench
People, places and events in charters: exploring the language of charters within ChartEx

Robin Sutherland-Harris
University of Toronto
Roger Evans
NLTG, University of Brighton
15 November 2013
Grant from Walter de Soke to Richard de Forde
Somerset Heritage Centre, DD\WHb/244
Sample document summary

Final concord made in the Court of Common Pleas at Westminster before James Dyer, Thomas Meade, Francis Wyndam, William Peryam, justices, in dispute between Richard Broke and George Armond, plaintiffs, and Robert Jenner, defendant, over lands (specified) in Fordham, Essex

Summary provided by the National Archives, Ward2_55A_188_30
Kanga Methodology

Fig. 1  The phases of the Kanga Methodology [36]. 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
http://dx.doi.org/10.1016/j.websem.2011.02.001
Initial Markup Schema: Guidelines

Basic Structure & Principles

ChartEx markup is directed towards highlighting and extracting information pertaining to particular geographical locations. As a result, certain aspects of the documents may be passed over – we are only looking for the components of each document that relate to our ultimate goal of reconstructing a topography of the medieval landscape. Thus, you may have to omit tempting and distracting details concerning, for example, the exact legal construction of the transaction in question. Remember that we are focussing on locations at all times – this can be a bit of a shift from the way historians typically conceptualise such transactions.

ChartEx markup is based around a number of entities, outlined below. These entities are tagged, and are subsequently linked to one another through relationships that also specify the role of the various entities in the document.

There are 4 main types of entities at work: **Actors, Sites, Events** and **Attributes**.

**Actors** includes **Persons, Institutions, and Actors**.

**Locations** includes **Sites** and **Places**, and is closely related to the concept of **Parcels**.

**Events** includes **Transactions, Dates, and Events**.

**Attributes** includes **Occupations**.

We also mark up a **Document** and **Apparatus** (any obvious editorial additions).

How to mark up and connect these entities is detailed below, but the basic principle is that they are defined in relation to one another, and in pursuit of information pertaining to locations.

It is also important to note that we want to be conservative in our interpretations of the document text. That is, our own suppositions and guesses about the wider historical context
Note that the Occupation of a Person may also appear at slight remove from the individual’s name. For example:
In witness to which: John, Simon, William, clerks of Bishop Joscelin...
Here John, Simon, and William all share the Occupation of clerk.

```xml
<owl:Class rdf:ID="Person">
  <rdfs:subClassOf rdf:resource="#Actor"/>
  <rdfs:comment rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
  <rdfs:subClassOf rdf:resource="#Site"/>
  <rdfs:comment rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:Class>

Most of the geographical units to be found in the documents we will be working with will be Sites. A Site is a specific geographical
Multiple Sites such as streets, fields, parks, cemeteries, schools, private properties can together make up Places (see below). Sites are
Some Sites may have proper names (as is more typical of Places) – streets, fields, parks, inns, religious buildings typically have proper
```
BRAT rapid annotation tool

ChartEx Architecture

Charter documents → Natural language processing → Analysed individual documents → Data mining → Analysed integrated documents → ChartEx workbench
ChartEx Architecture

Chart documents → Natural language processing → Analysed individual documents

Analysed integrated documents

Data mining

ChartEx workbench
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

[Diagram]

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.

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

(http://brat.nlplab.org/)
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.

... although actually to us, it looks more like this

(http://brat.nlplab.org/)
NLP development methodology

Step 1  find out what the ‘mark up’ should be (from historians)

Step 2  ask historians to mark up some documents manually

Step 3  use some of these examples to develop an NLP system that does ‘the same’

Step 4  use the rest to evaluate its performance (in progress)
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

**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.)
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.
Lexical layer: identify simple lexical phrases – groups of tokens that act as individual units. Promote other individual tokens to lexical items (with lexical types).
Lexical layer: identify simple lexical phrases – groups of tokens that act as individual units. Promote other individual tokens to lexical items (with lexical types).
**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.
Phrasal layer: use part-of-speech tags to build lexical items into local syntactic/semantic structures – and then ‘collapse’ all the equals relations.
**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
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;
How did we do?
Historians vs NLP

• NLP has to be told very clearly what it is looking for
• NLP has to make a lot of information explicit that people don’t notice
• NLP sees ambiguity where people do not
  – Really struggles with coordination and attachment
• NLP is programmed by computer scientists (or sometimes statisticians), not historians
• NLP can do it, but it definitely needs help!
Reconstructing spatial relationships from charters: a collaboration between Data Mining and Historical Topography

Stefania Merlo Perring, Sarah Rees Jones, York
Arno Knobbe, Leiden
This presentation will:

• Introduce the process used by Data Mining to reproduce the methodology of historic topography

• Discuss the results and further potentials

• Reflect on how Data Mining of charters and cartularies can add new meanings to Diplomatics
Matching Relational Information

Vicars Choral 408

Vicars Choral 409
Thomas son of Josce, goldsmith

- Statistics
  \[ p(\text{Thomas}) = 0.12 \quad (\text{common name}) \]
  \[ p(\text{Josce}) = 0.0015 \quad (\text{uncommon name}) \]
  \[ p(\text{Goldsmith}) = 0.04 \quad (\text{common profession}) \]

- Dating
  \text{vc-408} \ 1252-1253
  \text{vc-409} \ 1253-1261

- Final confidence
  \[ \text{conf} \ (\text{Thomas 408, Thomas 409}) = 0.9993 \]
Ranking of candidate Person-Person matches

0.99999  Alexander/Alexander, glover, witness
0.99998  Erard/Erard, witness
0.99998  Bernard/Erard, canons, witness
0.99998  Maurice/Maurice, canons, witness
0.99997  Adam/Ada, fisherman, sibling Richard
0.99997  Gilbert/Gilbert, chaplains, witness
0.99995  Roger/Roger, saddler, witness
0.99995  Roger/Roger, saddler
0.99994  Serlo/Serlo of Staingate, chaplains
...

THE UNIVERSITY of YORK
Thomas son of Josce matched
At the corner of Petergate and Stonegate
Connecting persons and sites between charters
Cartularies as networks

Charters of the Vicars Choral (selection of 124 charters)

confidence 0.99
Free from previous hierarchies

People, sites and events: all are nodes in a network
New indexes and catalogues
Conclusions/suggestions for debate

• People, institutions, things and technical devices represented as nodes (actors) in a network are freed from previous categories and hierarchies (Actor-Network-Theory, Callon 1991, Latour 1992, 2005)

• Nodes and patterns in a network are starting points for analysis and can generate new research questions and insights

• Networks connecting information from different fonds have potential for creating alternative virtual archives

• These can be created for a person or for a place (e.g. grouping charters concerning the same location)
The ChartEx Virtual Workbench

Helen Petrie, Chris Power, Dave Swallow

University of York
The ChartEx Virtual Workbench

A video presentation of the workbench is available on the ChartEx website at: http://www.chartex.org/docs/Chartex-Workbench-Demonstration-VIDEO.mov
Thank you!
ChartEx: Discovering spatial descriptions and relationships in medieval charters

www.chartex.org

@ChartExProject