People, places and events in charters: exploring the language of charters within ChartEx

Robin Sutherland-Harris
University of Toronto

Lynne Cahill
NLTG, University of Brighton
(now at University of Sussex)

4 July 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
BRAT rapid annotation tool

ChartEx architecture

Charter Documents

Data Mining

Analysed individual documents

Analysed integrated documents

Language processing

Researcher’s workbench
NLP in ChartEx
NLP Training Methodology

50 documents

Unmarked Documents

Newly marked documents

Training documents

NLP

Compare
What is NLP?

- **Natural Language Processing**
  - getting computers to process human language

- **Natural Language Understanding (NLU)**
  - aka Information Extraction
  - getting meaning from text
  - applications in wide variety of text types in modern languages

- **Ontologies used to help process texts and as target “meanings”**
NLP in ChartEx

- (Compared to much NLU)
- Short documents
- Standardised (formulaic) overall structure
- Complex referring expressions
- Phrases (especially noun phrases) often more important than clauses
- Usually key information within a single sentence:
  - Grant/quitclaim/enfeoffment etc.
  - by [person_A]
  - of [land_description]
  - to [person_B]
  - paying [payment]
Training data

• Three principle types of data provided by historians:
  1. Lists of key terms extracted from charters
  2. Raw (unannotated) charter documents
  3. Annotated charter documents
• 2 used to supplement 1
• 1 feeds directly into NLP lexicons
• 3 used for more complex analysis to train system and to test system against
BRAT markup

• Historians work with “pretty” interface
• For NLP training we use plain text output:

T1 Document 0 17 vicars-choral-387
T2 Transaction 18 23 Grant
T3 Person 27 34 William
T4 Institution 73 90 prebend of Masham
T5 Person 42 49 Richard
T6 Occupation 51 64 canon of York

... 

R3 is_son_of Arg1:T3 Arg2:T5
R4 is_grantor_in Arg1:T3 Arg2:T2
R5 occupation_is Arg1:T3 Arg2:T6
Differences

• Most NLP systems identify all (or at least many) possible interpretations, then rank them in order of likelihood

• Humans rarely aware of ambiguities a computer will spot:
  – *I met her yesterday*
  – Apparently unambiguous, but
  – *All our yesterdays*
  – Need to allow *yesterday* to be a noun
  – *So, met* has to have a person as its object?
  – *I met my deadlines*
  – Need to use general and contextual knowledge
A ChartEx example

Grant by William son of Richard, canon of York, to the prebend of Masham, for the salvation of the soul of the late king Richard and his own soul and the soul of his ancestors and parents of land in Petergate with buildings.

(Vicar’s Choral 387, first sentence)
Grant by William son of Richard, canon of York, to the prebend of Masham, for the salvation of the soul of the late king Richard and of his own soul and the souls of his ancestors and parents, of land in Petergate with buildings. The prebendary of Masham shall give each year on the day of William's death 20s. to the
Grant by William son of Richard, canon of York, to the prebend of Masham, for the salvation of
Grant by William, son of Richard, canon of York, to the prebend of Masham, for the salvation of the soul of the late king Richard and of his own soul and the souls of his ancestors and parents, of land in Petergate with buildings. The prebendary of Masham shall give each year on the day of William's death 20s. to the vicars of the church of St. Peter, but only to those present at his exequies. If the canon refuses to pay the rent to the vicars on the stated day, he shall be compelled to make satisfaction by the dean and chapter of York. If the canon is made dean or any other dignitary in the church of York, and having the houses does not wish to live in them nor give the vicars the said 20s., the vicars may lease the houses to whomever they please, and thus receive the 20s. Each prebendary of Masham at his first entry into the houses shall give the vicars 3 marks. Seal. Witnesses: [from text printed in YMF] R. the dean of York, W. the treasurer, tugistii Godard the penitentiary, J. Roman, T. de Lichefeld, Erard, Maurice, Robert de Winton, canons, mag. H. rector of the hospital of St. Peter of York, mag. Nicholas, mag. R. de Holbec, P. the
Grant by William son of Richard, canon of York, to the prebend of Masham, for the salvation of the soul of the late king Richard.
Evaluating NLP output

• Checking against manual annotation:
  – Does the NLP identify all the entities and relationships the manual annotation includes? (RECALL)
  – No, especially not relationships as yet
  – Are all the entities and relationships identified by the NLP in the manual annotation? (PRECISION)
    – No, but not all are incorrect

• Data mining can help distinguish important from irrelevant
Thank you!