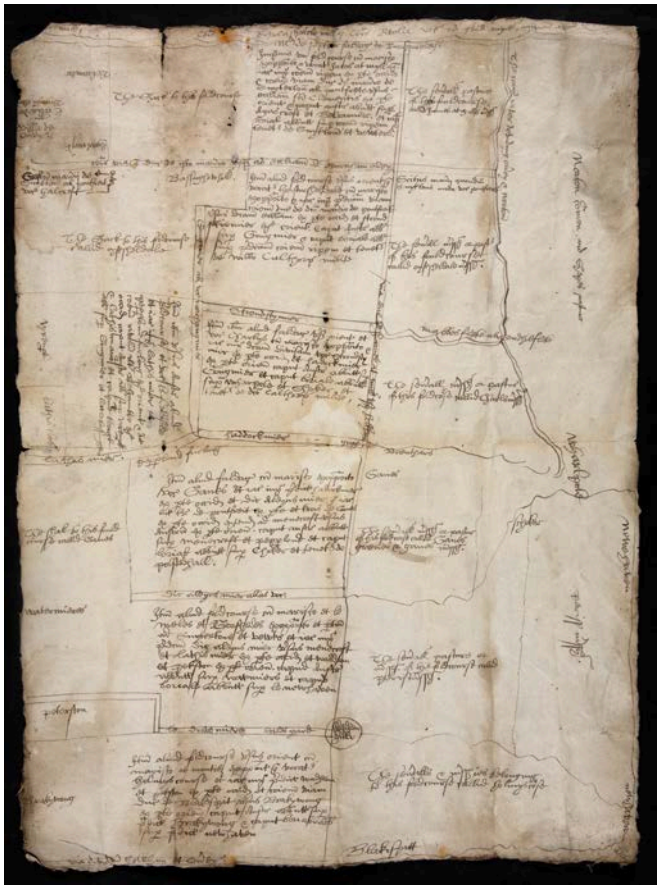
CHARTEX: DISCOVERING SPATIAL AND TEMPORAL DESCRIPTIONS AND RELATIONSHIPS IN MEDIEVAL CHARTERS.

WHITE PAPER 12 FEBRUARY 2014



WWW.CHARTEX.ORG @CHARTEXPROJECT

I. CONTENTS

1. LIST OF FIGURES

2. TABLES

II. INTRODUCTION

(Dr Sarah Rees Jones, University of York)

Researchers now have access to a deluge of data in the form of digitized historical records. One example are medieval charters[1] which record transfers of land ownership and are a major source for the study of people and places in the past, including the topography, economy and social relationships of pre-modern communities. However digital search aids are not sufficiently sensitive to the needs of researchers seeking to exploit the wealth of textual detail within this data. To make effective use of it researchers need better computationally-based systems.

The CharteEx Project aimed to research an innovative collection of computational methods to assist scholars in searching, analyzing, linking and thus understanding the content of charters. These methods could then be applied to other digitized texts, historical or contemporary.

ChartEx research focussed in particular on the extraction of information from charters, using a combination of natural language processing (NLP) and data mining (DM) to establish entities such as locations[2] and related actors, events and dates. The third crucial component of the ChartEx Project was the use of novel instrumental interaction techniques to design a virtual workbench (VWB) that will allow researchers to both refine the processing of the NLP and DM, and to directly manipulate (visualise, confirm, correct, refine, augment, hypothesize) relationships extracted from the set charters to gain new insights about the entities contained within.

The overall goals of the ChartEx Project were:

1. To develop and deploy a system that combines NLP and DM to extract data useful to researchers regarding locations and related actors, events and dates from digital charters
2. To investigate whether researchers working with a virtual workbench based on novel instrumental interaction techniques will produce more useful knowledge from charters than a human working alone or an automated NLP/DM system working alone
3. To investigate whether the ChartEx system can produce efficient and accurate knowledge from charter documents by processing the information in English summaries of the charters in comparison to the full Latin text of the charters

[1] Also referred to as title deeds, this paper will use the term charter for simplicity.

[2] Location in this context refers to a specific building or piece of land.

4. To investigate whether the rules used with Latin charters of UK provenance can be applied to Latin charters from different parts of Europe, or if not, how much modification of the rules is needed in order to produce accurate relationships.

A. DETAILS OF THE DATA CHOSEN FOR THE PROJECT

Charters are an abundant and fundamental source for the study of many aspects of medieval societies. While recent scholarship has expanded the range of charter studies to such fields as the history of emotions and performativity, their core usefulness of remains their provision of basic data: personal names, place names, and dates. In particular they help us to trace the ownership and occupation of houses and parcels of land over centuries providing the basis for many further studies from history to tourism and conservation. What makes charters so susceptible to Natural Language Processing is their relatively formulaic nature. The Latin (and after c.1300, vernacular) phrases that describe, for example, the location of a property (e.g., *'the tenement in Petergate lying between the tenement once held by John the apothecary and now held by Richard of Huntington on one side, and the church of St Michael on the other'*), were the pre-cursors to street-numbers and scientific spatial referencing developed from the 19th century. When researching a particular historical lived environment the researcher needs to establish links between actors, events, and locations, by recovering and reconstructing the relationships between hundreds, even thousands, of data points, included in different charters and even in different archives.

There are now many accessible high quality databases of marked-up charters that can support pioneering research into NLP and DM. In addition, there are thousands of older printed editions for which online text has been generated via OCR, and newer printed editions whose base texts exist in digital form. Finally archive offices have created many digital search aids to their collections, often including extensive summaries or full-text transcriptions of original charters. As a consequence the standards of digital data employed are diverse.

For the purposes of the ChartEx project we selected five core datasets including two in Latin and three in English:

1. LATIN RESOURCES

The DEEDS database (University of Toronto): over 10,000 digitised charters from the English Middle Ages before 1309. The charters are transcribed verbatim in their original language (Latin). Digitised using OCR and freely accessible both online via web services and locally via

DEEDS system[3]. For the purposes of the project we selected a subset of .charters relating to the county of Essex.

The CBMA (Chartae Burgundiae Medii Aevi) is a federated search that includes documents originally published in *Recueil des chartes de l'abbaye de Cluny*, ed. A. Bernard and A. Bruel, 6 vols., Paris (1876-1903). For the ChartEx project the digital texts were taken from the cartae cluniacenses electronicae (with permission): http://www.uni-muenster.de/Fruehmittelalter/Projekte/Cluny/CCE/Welcome.htm. The original charters are transcribed verbatim in their original language (Latin), providing the opportunity to test the ChartEx system against a different form of Latin from a different region (Burgundy) and period (11th century) from the DEEDS materials.

2. ENGLISH RESOURCES

The National Archives (TNA, UK), Court of Wards and Liveries: Deeds and Evidences (Ward 2): approximately 7,000 records from 12th to 17th centuries currently being provided with digital summaries in English and mounted online for public access[4]. A subset of the collection is currently available to the public using web services. For the purposes of ChartEx data was made available in EAD XML. We further selected a subset of the materials relating to the county of Essex.

Borthwick Institute for Archives, University of York: Catalogues for four deeds' series are available in digital format (OCR). We selected the catalogue for the Yarburgh collection, which is available to the public online[5]. This was made available to the project in EAD XML

Charters of the Vicars Choral of York Minster: City of York and its Suburbs to 1546, ed. Nigel J. Tringham (Yorkshire Archaeological Society Record Series 1993). With the permission of the author, and YAS, this was converted using OCR into machine readable text. Contains charters in full text Latin, full English translation and abridged English translation.

The datasets chosen for the project enabled us to develop solutions both for charters that had been transcribed in full in their original language and those that had been made available in English including those available from pubic archives online in highly abridged format in English. This latter solution is typical of the digital materials provided as a public service by public archives in the UK, while the Latin resources, although made available to the public online, were typically designed to support scholarly research by expert academic researchers.

[3] http://www.utoronto.ca/deeds/

[4] http://www.nationalarchives.gov.uk

[5] http://www.nationalarchives.gov.uk/a2a/

ChartEx quickly encountered two problems with the English resources. Of these the most significant was the lack of a standard approach to either the abridgment or the digitisation of the material even though both archival collections used the standard markup language (Encoded Archival Description). There were radical differences in the design and provision of both metadata and data content. The materials from TNA were so abbreviated that most of the details of parcels of land were omitted (save the name of the town or village in which it was located). For ChartEx some of these details were restored to a sample of the material. In the case of the Borthwick the decision to include key entities such as dates and titles of transactions in the metadata, but not in the data content, caused problems in the historical interpretation and thus the training markup of the texts. The two collections of Latin charters did not employ any standardised metadata system, each had developed their own, but since both provided full text transcriptions of the data content this was not a problem.

To solve these problems, at least in the initial stages of the project while developing a preliminary ontology, we therefore created a new dataset, derived from a publication which included both full text Latin and full text English from a total of c. 600 charters. In this case (the Charters of the Vicars Choral of York) the main problem was the low quality of the OCR digital transformation of the text which introduced many errors that need to be manually cleaned.

From problems with the digital two recommendations to archives emerges:

Recommendation 1: Full text transcription and/or translation of archival texts is preferable to abridgement of texts. Abridgement is particularly problematic due to the lack of a standard approach.

Recommendation 2: The application of Encoded Archival Description in the provision of archival metadata is highly idiosyncratic, not only between archives but even between individual archivists. The major problem is that in some instances this creates a barrier to the interpretation of the meaning of the original record.

III. RESULTS OF THE CHARTEX RESEARCH

A. DEVELOPMENT OF THE CHARTEX ONTOLOGY

(Robin Sutherland-Harris, University of Toronto)

A fairly basic but useful definition of an ontology, can be found in the work of Natalya Noy and Deborah McGuinness. They define an ontology as "a formal explicit description of concepts in a domain of discourse... with properties of each concept describing various features and attributes of the concept."[6] While this definition was written with computer scientists in mind, within the ChartEx Project we first focussed on establishing a set of useable criteria for the information we would generate that was both accessible to historical scholars useful for programmers. Thus in the end, our markup guidelines can be considered simply as a different expression of our rdf schema. We found this ontology development to be a useful process for ensuring an explicit, defensible, and sharable articulation of underlying assumptions and vocabularies. We believe, as do many other scholars involved in digital projects ranging from archaeology to cultural heritage to history, that such ontologies can and should be a transparent and public part of digital humanities projects rather than an embedded, behind-the-scenes component.[7]

Within the ChartEx project, an ontology was a necessary step to mediate between the expertise of our charter historians and the technological processes of NLP and DM. We first had to train the NLP software by feeding it a body of texts that modelled the kind of output we wanted to achieve. This meant manual markup of over 200 charters or charter summaries. It was clear that traditional scholarly approaches to these materials would not help us address our main question of topological reconstruction, mainly because of differences in focus and goals, but also because many of the materials used in ChartEx are summaries of charter material, and so do not contain sufficient formulae or original text to allow for traditional approaches. Our mark up schema was generated in response to two factors - first, and most obviously, the particular goals of the ChartEx project, which steered us away from some already well-established frameworks, and second, the requirements of the technology we use in arriving at this goal.

1. PROCESS

The process of developing our ontology was straightforward. Because ontology development is a relatively new field, and one that operates in a vast array of environments, there is no established approach for how to go about producing one. Initial ideas drew on already

[6] http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf, p3.

[7] Jeremy Hugget, "Lost in information? Ways of knowing and modes of representation in e-archaeology," *World Archaeology* 44.4 (2012), 544.

established ontologies, in particular that used in the mark-up of legal formulae in similar documents at the DEEDS database.[8] This provided some valuable early direction and established a basis for refinement and adaptation towards the goals of the ChartEx project. For this, we used a methodology specifically developed to give domain experts - in our case our charter historians, control over ontology creation.[9] This method, called KANGA, begins with defining the purpose and scope of the ontology, thereby setting some boundaries to restrain its expansion, and by listing already available resources and materials. Then the charter historians collectively worked through a number of documents to highlight the entities and connections between them that would be necessary in getting us manually from a group of texts to a reconstructed historical topography. These entities and connections were then used to map out a draft schema, which we tested and refined, and which our NLP and DM partners commented on. This process was repeated several times, until a complete schema and explicit, written guidelines for its application could be settled on. This method had as a particular strength the fact that the ontology was generated by scholars deeply familiar with the types of source materials being used and with the kinds of issues likely to arise, and it is also helpful that the ChartEx historians exemplify some of the target end-users of the project.

Several challenges arose during the process of developing our ontology. The first concerned scholarly habits and the nevitable influence of longstanding methodologies for considering these types of historical records. Early versions of the schema contained redundant or irrelevant information and tended to take an approach more in line with traditional scholarship. For example, in traditional scholarship, the dispositive clause of a charter is crucial in identifying what type of document is being considered (quitclaim, grant, inspeximus, etc.). However, although we all had the first instinct to highlight this information as significant, we soon realised that in fact, the type of document at hand is immaterial to the extraction of topographical information; all that was really needed was the simple fact that a transaction of some kind occurred to connect actors and locations. That said, it's important to note that the accurate and effective markup of documents according to the ChartEx schema would not be possible without researchers familiar with traditional categories of analysis applicable to such sources. In fact, the documents in question (in particular some of the truncated summaries of charters) are often formulaic in such highly particular ways that it can be difficult to understand fully what they mean without a thorough appreciation and understanding of charter formulae and

[8] The DEEDS database provides a valuable source of already digitised medieval Latin charters, and makes use of well-defined systems of mark-up designed for specific purposes. Its usefulness to the ChartEx project rests not only in providing a jumping-off point for ontology development, but also in serving as a rich source of Latin materials available for training and testing the NLP and Data Mining components of the project.

[9] For an overview see Ronald Denaux, Catherine Dolbear, Glen Hart, et al., "Supporting Domain Experts to Construct Conceptual Ontologies: A Holistic Approach," *Journal of Web Semantics* 9 (2011): 1-23.

the scholarly tradition that surrounds them. A more broadly applicable challenge is that category definition via mark up schemata and ontologies shapes knowledge, which can be an advantage in seeking answers to particular questions, but should not be overlooked in later stages of project development. As Jeremy Hugget has pointed out, standards and standardization are not neutral, either within an individual project or within a larger scholarly community.[10] These challenges reinforce the argument for public and transparent ontologies in digital research projects.

2. RESULTS

This process resulted in a schema that has three main entities (with a few sub-types): locations, actors, and events (including dates). In keeping with the importance of understanding and unpacking quite complex relationships between people and places that is required for reconstituting medieval topographies, there is a comparatively large emphasis on relationships between entities. We have 26 categories of relationships, and many more actual connections that can be drawn between entities. This was written up as both a set of guidelines for historians engaged in document markup and as an rdf schema. The manual markup of our training charters for natural language processing was completed using standoff markup via an online tool called BRAT (Brat Rapid Annotation Tool), which can express the markup as rdf triples. The clarity of our ontology at this stage allowed for moderation or "proof-reading" of the markup and thus a reasonably high level of consistency in its application.

3. IMPLICATIONS

The markup that we've developed has a different focus from other charter scholarship, especially that which already has digital form. While this is to be expected given the nature of our project, the deeper implications of digital approaches to scholarship are also particularly evident here. We've been forced to step aside from older categories and patterns of analysis and articulate some alternate ones, though these too have roots in traditional analogue practices. The study of charters is largely conditioned by scribal practice, and logically therefore focusses on the text itself. In contrast, the ontology developed for ChartEx, quite apart from the purpose it serves in the project itself, is conditioned by the events, or transactions, which

[10] Hugget, 543.

relate individuals, institutions, and locations to one another, and thereby give rise to the creation of the text we read. This also raises the possibility of alternative perspectives: questions could be posed of data structured in this way that foreground either or locations or actors, or focus on particular types of relationships.[11]

The larger point here is that digital projects with their peculiar questions and technical demands, even before they get anywhere near completion, in and of themselves can constitute a process by which scholarly categories are broken down, reconsidered, supplemented, or reified, even when that is not at all the goal. It is wise to be forthright and transparent about this process, both to enable critical discussion and to remain attentive to intriguing new avenues of approach that might otherwise be brushed aside.

B. NATURAL LANGUAGE PROCESSING IN CHARTEX

(Roger Evans & Lynne Cahill, University of Brighton)

1. INTRODUCTION

In this section we discuss the natural language processing (NLP) component of the ChartEx system. Broadly speaking the NLP component can be characterised as a software application which takes individual charter documents and attempts to extract key information about people, places and transactions from them. This semantic information, represented in symbolic form, is passed to the Data Mining component, which analyses semantic information coming from many documents, to identify correspondences between documents and generalisation across documents. Thus the NLP component is responsible for all the 'linguistic' processing in ChartEx, but only considers documents in isolation.

In this section we discuss the overall context for the NLP development, including an overview of the manual linguistic analysis which supported it (discussed in more detail in section NN above). We then summarise our overall technical approach, including the key linguistic and computational challenges of the task, and the architecture of our solution. ChartEx uses an innovative NLP architecture, based on an 'extended' notion of lexical description implemented using default inheritance-based techniques. ChartEx is the first significant demonstration of the

[11] Some relevant discussion can be found in Michael Ashley et al., "Last House on the Hill: Digitally Remediating Data and Media for Preservation and Access" ACM Journal of Computing and Cultural Heritage 4.4 (2011): 13:1-13:26.

effectiveness of this approach. We then provide a walkthough of the main stages in NLP processing and finally discuss evaluation of the system in terms of meeting both technical and project goals.

2. TECHNICAL CONTEXT OF NLP DEVELOPMENT

The ChartEx project involved teams of technical specialists, focusing on the NLP, Data Mining (DM) Expert Elicitation and Virtual Workbench (VWB) and system architecture aspects of the system, together with teams of historians, each with specific expertise in different aspects and collections of charters. The overall structure of the technical work in the project is shown in figure 1.

In the first phase of development a very small number (5-10) of charters, carefully selected for their representativeness, were used to undertake detailed elicitation from the historians. Through in-depth workshops and interviews (described more fully in section III.A.1), discussing and observing exactly how historians work with charter documents and for what purposes, two technical aims were achieved. On the one hand, we developed a set of requirements for the VWB, setting out exactly what kind of facilities and interfaces were best suited to supporting historians at work. On the other hand we developed a markup scheme, consisting of a symbolic language to represent the semantic information we needed to extract from charter documents to support the VWB, and a comprehensive set of guidelines explaining how this language could be used to represent information expressed in charter texts (giving guidance, for example, on how to decide when to treat a ‘dean’ as a person and when as an institution).

The development of the markup scheme was a significant milestone in the project, not only because it gave us a concrete language to act as the foundation for all the technical work, but also because it forced us to come to a common understanding of what that language would be – of what, for the purpose of the ChartEx system, charters actually meant. This was a significant challenge, given the range of disciplines involved, and a vital step in the progress and success of the project.

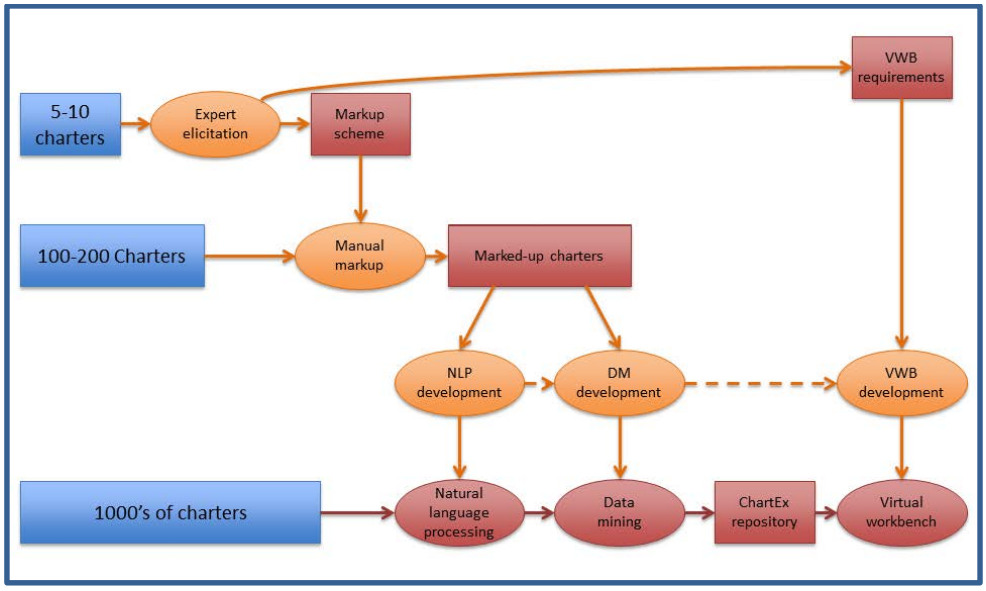


Figure 1: Overview of the ChartEx development structure

The second phase of development involved taking a larger number of charters, but still only a couple of hundred, and analysing them manually, according to the markup scheme devised. This analysis is a task requiring detailed expertise of charters, and so had to be undertaken by the historians in the project. The approach taken was to analyse a charter by adding 'annotations' to it, some identifying individual text strings as entities (such as persons or places), and some identifying relationships between entities (such as ownership or transfer of property). To do this we used the BRAT annotation tool (Stenetorp et al 2012), which provided a user-friendly web-based interface for adding and visualising annotations. Figure 2 shows an example of the BRAT interface, displaying a range of semantic relationships as annotations on a charter text. So for example, 'Thomas' and 'Josce' are identified as of type 'Person', and there is an 'is_son_of' relationship between them. In this second phase charters were manually annotated, moderated and cross-validated (by subjecting a selection of charters to analysis by different historians and checking for inter-annotator agreement). This set of charters was then available to support the development of both the NLP and DM processing components.

The third phase of development was the core development of the processing tools themselves. These tools are intended to operate more or less in a pipeline, with NLP processing feeding the DM, which populates a repository which is used by the VWB. However within the relatively short timescale of the project, a pipeline structure for development was not feasible. We took the decision to use the BRAT annotation language as the language for communication between the NLP and DM components. We had originally envisioned some variant of RDF-XML for this interface, and while formally there is a fairly straightforward correspondence between the two, the practical advantages of using BRAT were that the DM development could make use of the manually annotated charters (originally intended only for NLP development) directly, evaluation of the VWB could proceed earlier in the project, and we had a convenient tool to visualise NLP component output.

Figure 2: The BRAT annotation tool

The final stage of development is running the entire pipeline over a substantial body of charters. At the time of writing this stage has not been fully achieved: although all the links in the pipeline have now been tested, a full scale run has not yet been undertaken.

3. NLP REQUIREMENTS

Arising from this architectural description, we get a clear idea of the requirements for the NLP processing component. Starting with a charter text such as the one shown in figure 3, our goal is to produce a network of semantic relationships for the document, such as the fragment shown in figure 4.

> 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.

Figure 3: Charter 408 from the Vicars Choral collection (English summary).

Given our decision to use BRAT as our semantic representation language, the way we actually think of this network is in terms of annotations on the text, as show in in figure 5. But even this is just a visual representation of an underlying textual form – BRAT standoff annotation – as shown in figure 6.

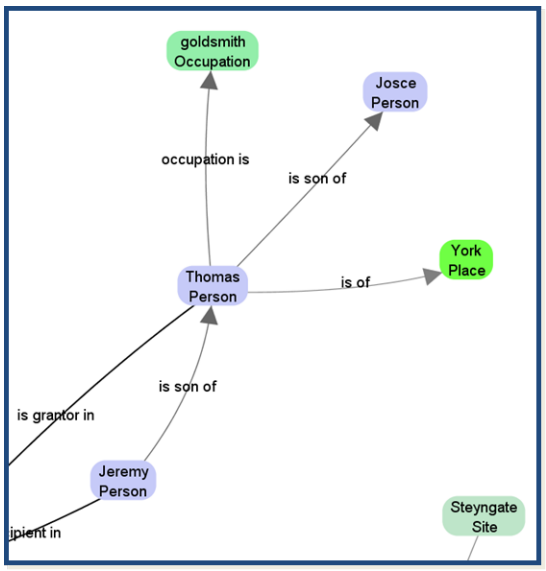


Figure 4: Semantic relations extracted from text.

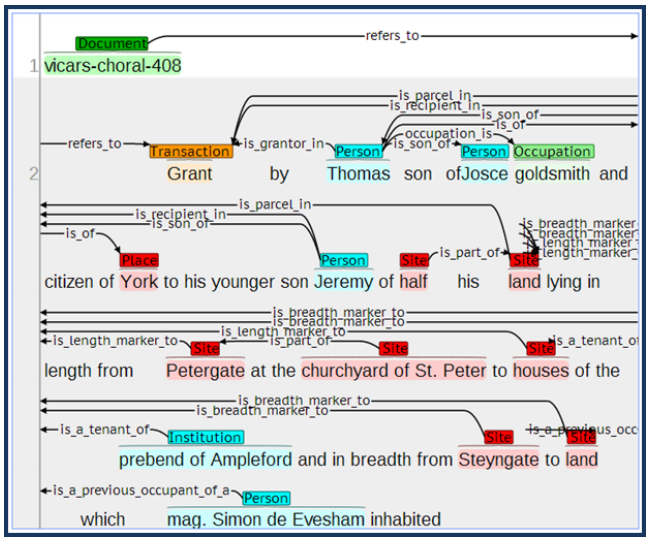


Figure 5: Semantic relations as BRAT annotations.

In figure 6, each line corresponds to an individual entity or relationship. It has an identifier (T1, T2, R1, R2 etc.) and a type (Document, Person, is_son_of etc.). For entities, the rest of the line simply identifies the part of the text corresponding to the entity (in character positions in the document, and the text itself), while for relations, it specifies the entities involved in the relationship.

The upshot is that the NLP component implements a mapping from text input documents like figure 3 to text output documents like figure 6.

One further note about the context for the NLP development is relevant here. The number of charters available for development is by NLP standards quite small. Indeed a state-of-the-art statistical NLP system might require several thousand documents for effective training. However providing that much manually annotated data is a significant practical challenge for a real application domain, especially a relatively specialised on such as ChartEx. The basis of our approach to ChartEx was that charter documents are sufficiently constrained to allow us to use primarily symbolic NLP methods, which are in general example based, and make more intensive use of a smaller set of examples for development. Indeed one of our key longer term goals with this work has been to investigate how symbolic NLP methods can be used to achieve some of the coverage advantages generally claimed for statistical systems.

```
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
```

Figure 6: BRAT stand-off annotation 'under the hood'.

4. TECHNICAL PROGRESS

A) CHARTERS – THE LINGUISTIC CHALLENGE

Charters are available from roughly the period between 1200 and 1600. They are written in either English or Latin (often a mixture of the two) in a fairly standardised form of ‘legalese’. The first phase of the NLP development focused on the English charters, specifically the Vicars’ Choral Collection (VCC). The VCC has been transcribed and (in the case of those written in Latin) partially translated and published in book form. This book has been converted to PDF and OCR has been used to digitise the documents.

The challenges for an NLP component are both technical and linguistic. As with any document which has been processed using OCR, there are errors. In the case of these charters, this particularly applies to numbers and certain letter combinations which are easily confused. The

OCR often gets word spacing wrong so that "son of Josce" comes out as "son ofJosce". Some of these errors can be relatively easily dealt with automatically, but others cannot.

The linguistic challenges are more interesting. The language used is formulaic, and the overall structure of the charters follows a fairly consistent pattern. Each charter is on average around 150 words long, and broadly conforms to the following structure:

[TRANSACTION] {by, from} [PERSON1] of [PARCEL] to [PERSON2] paying [PAYMENT]

where TRANSACTION is one of a restricted set of terms describing a legal transaction such as *grant, quitclaim* or *enfeoffment*; PERSON1 and PERSON2 are referring expressions identifying the person or persons involved in the transaction, PARCEL is a referring expression identifying the property involved in the transaction and PAYMENT is the payment involved. These three referring expressions, however, can be extremely long and complex. For example, the following are all examples of referring expressions identifying people:

- Thomas son of Josce goldsmith and citizen of York
- Ellis de Sutton clerk to Roger de Wyghton and his wife Margery
- Simon de Botelesford warden of the house of vicars of the church of York and the vicars
- Margaret widow of John Damysell of York coteller
- mag. Geoffrey de Norwich (Norwico) once precentor of York

A cursory glance tells us that there are a lot of potential ambiguities and uncertainties in these expressions. Which of Thomas and Josce is the goldsmith and/or the citizen of York? Is Margery the wife of Ellis de Sutton or of Roger de Wyghton? Some of these issues can only be addressed with significant background knowledge of the area and the people and places involved in earlier transactions, and this is the knowledge that is brought to bear by the historians who study the charters. Some is a matter of convention and understanding of the norms and institutions of the time (a woman at that time would be unlikely to have a clerk, for example, and it would be more likely that the occupation of the person involved in the transaction would be provided than the occupation of their father).

There are also issues relating to how to classify the entities described by phrases like "*the house of vicars of the church of York*". The "*church of York*" could be referring to the building, and so classified as a site, or it could be an institution. Similarly, "*vicars of the church of York*" could be referring to a group of individuals or to an institution and, in practice, our historians annotate the whole phrase as an institution, with no reference to the sub-phrases within.

B) THE DATA ANALYSIS

In order to develop an NLP component which could interpret these expressions we needed to incorporate the historians' knowledge of how to interpret the language. As discussed above, we took a sample of charters and asked a panel of historians involved in the ChartEx project to

manually annotate them, indicating the key people and places involved in the transactions, using the BRAT annotation tool (Stenetorp et al., 2012). The resultant annotations can be visualised as markup on the charters themselves and this allows us to immediately see the stretch of text which provides the specification of the entities involved, although the same is not true of the relationships, for which no anchor in the text is given (the relationships are marked on the links between the entities which are anchored in the text). Entities can be people, places or events, although we are not currently using any events.

The data analysis took a combined manual and automatic approach. We began with a set of 50 charters from the VCC and Ward2 sets, all of which were in English, and manually examined them to identify the common structure and the kinds of linguistic expressions used. We then took the text spans from the annotation files for each entity type (person, site, transaction etc.) and examined them to identify specific patterns for each entity type. Finally we extracted vocabulary using the SketchEngine (Kilgarriff et al., 2004) corpus analysis tools to give us lists of terms used in similar ways to include in the lexicon, and to identify patterns of sub-categorisation and collocation.

The analysis process resulted in two different types of data which we incorporated into the system. The lexical resources came from a combination of lists of terms which the historians had previously collated and additions to those lists identified by the SketchEngine. The phrasal analysis led to a set of phrasal rules which specified the structure of the various referring expressions.

C) THE NLP SYSTEM

The NLP system in ChartEx is an application of the 'Extended Lexicon Framework' (ELF) introduced in (Evans 2103), which in turn builds on the default inheritance-based lexicon description language, DATR (Evans and Gazdar 1996). DATR is a knowledge representation language designed for capturing lexical information using default inheritance. It has been used particularly to model morphology, phonology and syntax, but on its own it only models individual lexical entries, that is, words in isolation. The assumption is that some other language component (such as a parser) makes use of the information provided by a DATR lexical database.

ELF uses DATR to model not just words in isolation, but words in context, that is, words within a sentence. At the most basic level, a word in context differs from an isolated word solely by knowing what its neighbouring words are. More specifically, it can access information about its neighbouring words and use this to condition its own behavior. As a simple example, the node for the word 'a' could look to see if the following word starts with a vowel, and if so change its own form from *a* to *an*.

This simple idea allows us to combine the descriptive power of DATR's default inheritance with the ability to undertake 'non-lexical' linguistic processing. In (Evans 2013) we describe how to

undertake part-of-speech (POS) tagging in ELF: each node has a feature 'pos' which is in terms of the 'pos' values of the preceding 2 or 3 words (a word instance only has direct access to its own neighbours, but it can ask a neighbour about its neighbours and so access information from further away in the sentence). When an actual sentence is instantiated, the values for the 'pos' feature on each instance node is determined by inspecting the pos values of previous nodes and using an (inherited) POS tagging model to determine the correct value. For example, the 'pos' value for *saw* might be *noun* if preceded by a word whose 'pos' value is *determiner* or *adjective*, but *verb* otherwise.

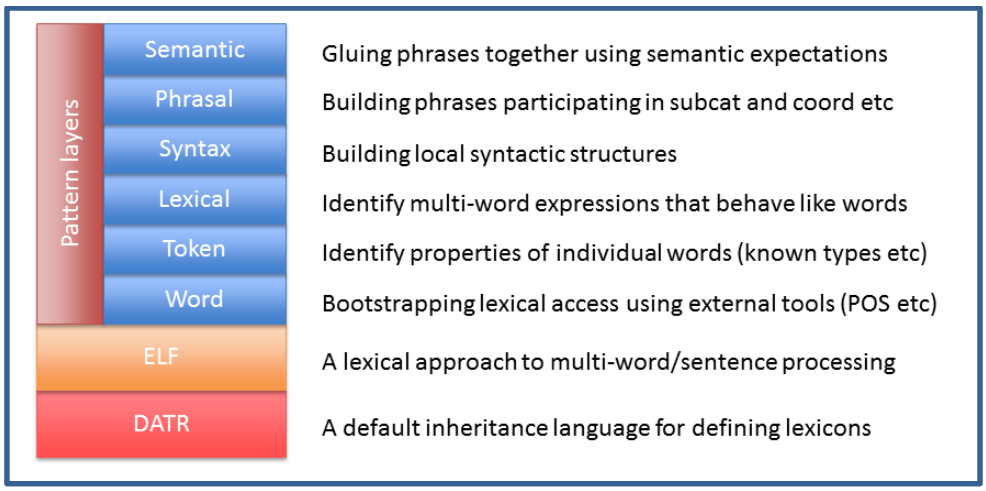


Figure 7: NLP system architecture

In ChartEx ,we have built a large scale semantic analysis engine using ELF. The overall architecture of the system is shown in figure 7. On top of DATR and ELF there is a collection of layers of processing, from individual words up to semantic constructions. Each layer sees the text as a sequence of items, and adopts an ELF-based view of processing those items – each item can have information associated with it directly, or by virtue of information obtained from its immediate neighbours. The general model is that each layer uses finite-state pattern matching over the layer below to generate new items. The effect is similar to a cascade of deterministic finite state transducers, except that DATR's descriptive power allows more complex operations, and the specification of the transducers can exploit DATR's default inheritance to combine powerful generalisations with subclasses and exceptional behaviour associated with specific words.

The overall aim of the system is a quite traditional information extraction task, processing each individual document and producing BRAT annotation output for it. For example, a document fragment such as *"William son of Richard, canon of York"* produces the following BRAT annotation output:

T3 Person 27 34 William

#4 Property T3 gender = male

T4 Person 35 49 son of Richard

#5 Property T4 gender = male

T5 Person 42 49 Richard

#6 Property T5 gender = male

R3 is_son_of Arg1:T4 Arg2:T5

T6 Person 51 64 canon of York

#7 Property T6 occupation = canon

T7 Occupation 51 64 canon of York

R4 occupation_is Arg1:T6 Arg2:T7

R2 same_as Arg1:T4 Arg2:T6

R1 same_as Arg1:T3 Arg2:T4

This asserts that the string “William” refers to a male person, the string “Richard” also refers to a male person, the strings “son of Richard” and “canon of York” also refer to people, the latter having occupation “canon”, “canon of York” also describes an occupation, and finally that William and the canon of York are the same entity, and the canon of York and the son of Richard are the same entity.

The core of the system is an ELF lexicon which contains lexical knowledge of the following sorts:

- specific words and word classes with specific hand-crafted behaviours, either because they are syntagmatic (eg coordinators) or because they relate directly to the extraction task (such as familial relation syntax).
- individual words and phrases extracted from the training charters and elicited from expert historians, for example, lists of occupations and institution types
- gazetteer information: places, institutions (particularly religious institutions), personal names
- Handling for otherwise unknown words.

Our overall aim has been to balance the amount of handcrafted knowledge against a dependence on specific empirical data associated with the training corpus or background gazetteers.

Our approach to semantic analysis has two opposing facets. On the one hand we specify how linguistic analysis are constructed from the bottom up, starting with words, then tokens, then lexical phrases, local syntactic phrases and semantic constructions (as shown in the next

section). On the other hand, the actual analysis process is goal driven from the top down. The system sets out to identify a transaction, and then likely participants and properties in it, then relationships between people or describing properties. It completely ignores sections of text that do not contribute to these goals, however, making it quite robust in the face of unexpected information.

D) EXAMPLE WALKTHROUGH

In this section we illustrate the key steps in the NLP analysis of a charter. As a running example, we will use the sentence shown in figure 8, the first sentence of charter VCC 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;

Figure 8: The first sentence of charter VCC408

Processing a document in ELF has two stages. First, the document has to be incorporated into the ELF lexicon, by linking each word in the document to an appropriate abstract lexical entry, and creating links between them so they can correctly access their neighbours. Once incorporated, analysis proceeds simply by querying word entries – typically asking the last word to provide an analysis of the whole document. For the first stage, we use some external tools, including a tokeniser and (simplified) part-of-speech tagger provided by the openNLP[12] package and some additional information directly derived from the word token, including the word form itself, the word normalised to lowercase, a simple morphological analysis (stem and morphological features), and its 'case type' (lowercase, uppercase, capitalized, mixed). The result is shown in figure 9.

[12] http://opennlp.apache.org/

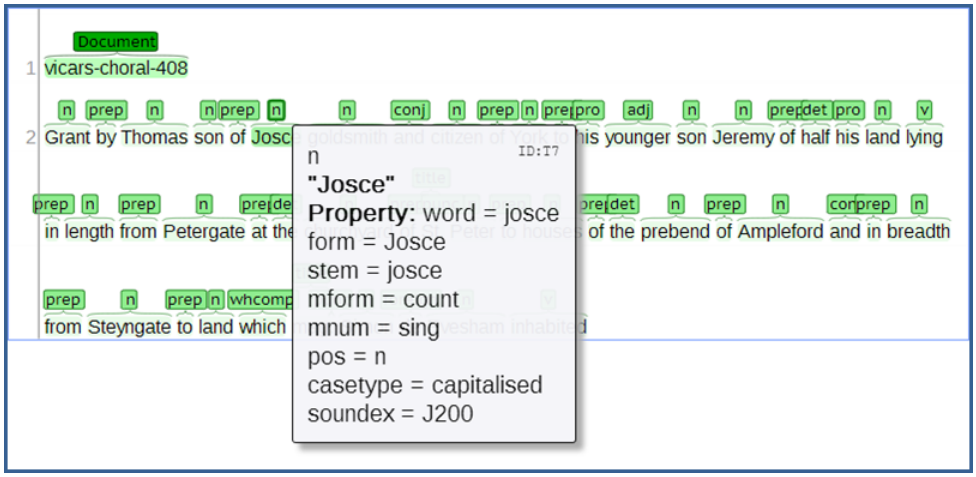


Figure 9: The word layer: basic information about individual words, collected 'outside' the main ChartEx NLP system

These word-level features drive the indexing into the ELF lexicon. ELF first looks to see whether the exact word form, or the normalized word, or the stem, has an explicit definition of its own, and if so it uses that. Use of exact word forms (including case distinctions) is rare, but an example might be a title such as *"King"* or *"Saint"*. The normalized form is useful where the exact morphological form is required, typically as a component of a fixed phrase. The stem form is the most common, where one abstract node supports multiple forms of a word. For example in the occupation *"bridge keeper",* the first word should link to the normalized form while the second links to the stem. This ensures that *"Bridge keeper"* and *"bridge keepers"* are recognised, but *"bridges keeper"* is not.

If no explicit definition associated with the token is found, then lists of known word types, derived from corpus analysis are consulted, to establish if the token is, for example, a firstname, surname, placename, sitename, occupation or institution. If so, it is assumed to be a regular instance of its type, and linked to a generic abstract node for that type.

If this step fails, the word is not directly known to the system. At this point we use the part of speech tag to establish a possible general syntactic role of the word, allowing it to participate in phrase building with known words. This is key to building up larger phrases without a very complete lexicon. Note that the part of speech tags are derived from the external openNLP POS tagger, so we are implicitly dependent on its coverage here. Finally, if the POS tag cannot help us, the token is linked to a generic 'unknown word' abstract node, which allows it to participate passively in other constructions if required.

Once a word has been successfully linked, it becomes a token, and may have type and other intrinsic information, such as gender, associated with it, as shown in Figure 10.

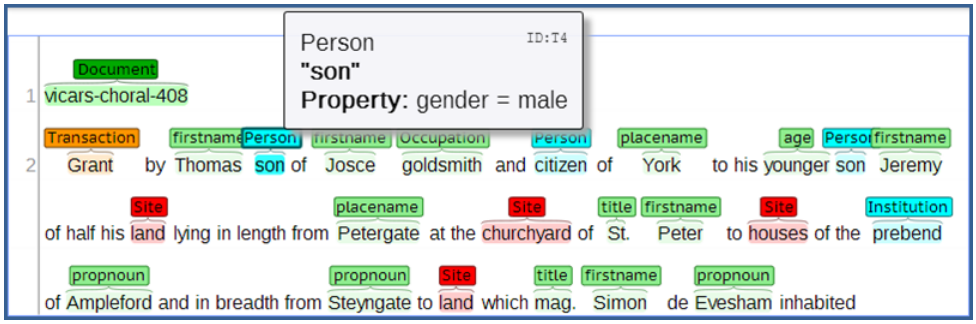


Figure 10: The token layer: identify semantic types and intrinsic properties (eg gender) of known individual words

Once the token nodes have been attached to the abstract lexicon, the links between neighbouring modes are establish, and the analysis process can be undertaken. As discussed above, analysis tasks are conceived simply as part of lexical lookup. A document is 'processed' by dynamically incorporating it into the ELF lexicon and then asking the first word to return the analysis result (in the form of a BRAT annotation as above). It is the responsibility of the first word to interrogate the rest of the document as necessary to determine this result. And of course, what it returns will in general vary according to the other words in the sentence (although the abstract nodes they link to remain the same). All the work of analysis is done by words interrogating their own content and that of their neighbours and conditioning their own responses on the basis of what they find.

The actual processing architecture is layered. The incorporation process described above establishes the **token** layer, in which each lexical token is a separate item linked to its abstract word node as described above and to its neighbours. Above this sits the **lexical** layer (See figure 11). By default, the lexical layer inherits from the token layer, if you ask the lexical layer for information about a word, it will just return whatever is specified at the token layer. However, the lexicon layer can override this default in various ways. In particular the lexicon layer is responsible for building lexical phrases – multi-word expressions that are considered to be syntactically and semantically atomic.

An example of this is the phrase *"prebend of Ampleford"*. At the token level, there are three tokens here, and the first of these, *prebend*, will be identified as an occupation word and so linked to the general abstract occupation type. However, if you ask *prebend* about its lexical type, the abstract occupation type node will look to see if the next token is *of* and if so, if the one after that is a place name (either a known place name, or failing that at least a capitalised word). If so, *prebend* absorbs the entire phrase into a single lexical unit.

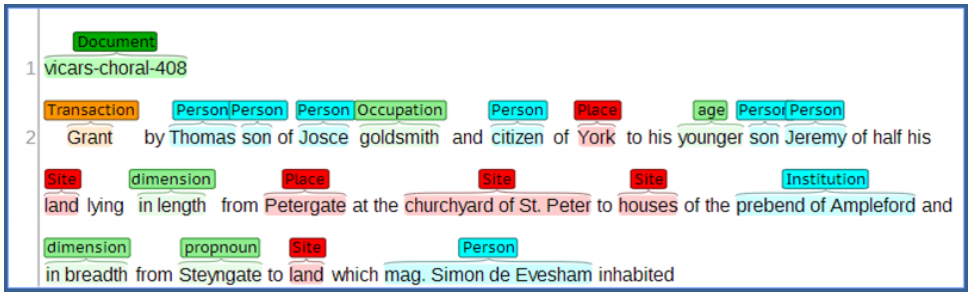


Figure 11: The lexical layer: identify simple lexical phrases – groups of tokens that act as individual units.

Each layer has its own notion of previous and next item built on the layer below it. At the token layer, previous and next correspond directly to the tokens in the input document. At the lexical layer, lexical phrase processing can intervene, so that, a phrasal unit is considered a single item. So if one asks *prebend* what the next lexical item is, it will return information about the item after *"prebend of Ampleford".*

Above the lexical layer is the **syntax** layer see figure 12). This layer is responsible for combining lexical units into local syntactic phrases. Like the lexical layer, the syntactic layer has its own notion of previous and next phrase, and by default this uses the lexical layer itself (which by default use the token layer, so the simplest phrase is often just a word – *"Jeremy"* is a lexical noun-phrase). But the phrasal layer implements simple right-recursive phrase recognition to build up more complex phrases. Some of these are word, or word-class, specific, while others are generic.

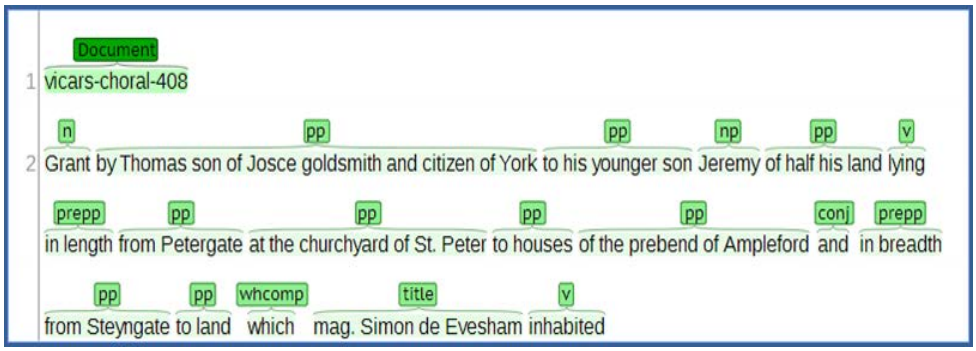


Figure 12: The syntax layer: build (local) syntactic structure to identify basic constituents of the sentence.

The **phrasal** layer combines this syntactic information with semantic types associated with tokens to build local semantic relationships. The phrase *"son of Josce"* is an example of word-class specific phrase building. *Son* is identified as one of a fixed set of family relationships which are of interest to the domain, and hence linked to a specific family-relationship abstract node. At the token and lexical levels, this does nothing, but at the phrasal level, it looks for a following

of token and then a phrasal noun-phrase. In this case, *"Josce"* is a lexical noun-phrase and hence also a phrasal noun-phrase (by default), so *son* is able to build a predicative modifier phrase covering *"son of Josce".* If *son* had not found *of* then it would have done nothing, and remained a simple noun. Note also the use of recursive phrase matching, which would allow more complex noun-phrases to be incorporated, such as *"son of the goldsmith".*

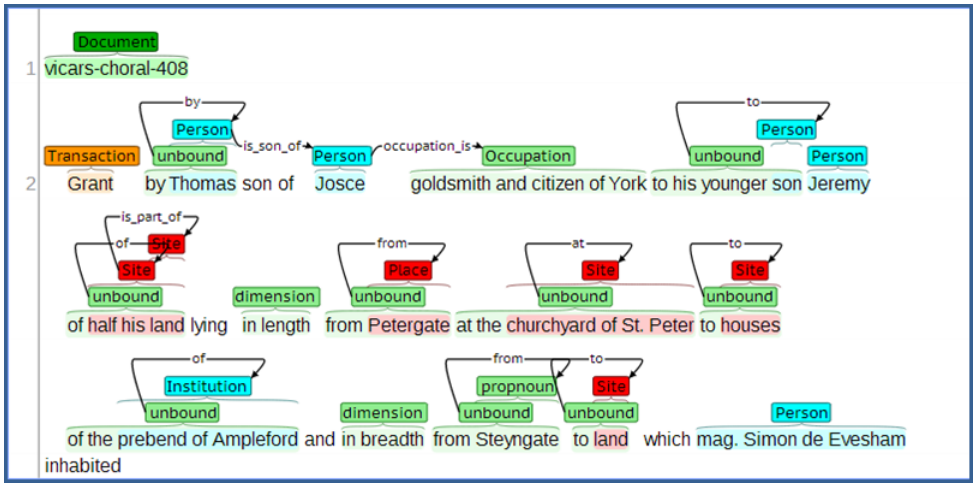


Figure 13: The phrasal layer: use part-of-speech tags to build lexical items into local syntactic/semantic structures. These have lots of unbound arguments – like jigsaw pieces waiting to be slotted together.

More generic phrase building includes typical examples such as noun-phrase and preposition-phrase building. Our approach is strictly left-to-right, and driven by the simplified part of speech tags (unless overridden by specific entries for individual words). Thus a determiner will trigger the search for possible adjective followed by a nominal phrase etc. In general, our analyses are fairly flat – the architecture gives us good control over individual patterns of syntax associated with particular words and phrases.

The phrasal level only considers local relationships, comprising words or phrases that are strictly adjacent to each other. Because of this, it tends to create small pieces of semantic knowledge with gaps in them, like jigsaw pieces waiting to be joined up. For example *"by Thomas"* represents a relationship between *Thomas* and something else (usually a transaction, in this case a grant). The phrasal layer creates the relationship, but does not fill it in – in general the filler might not be adjacent to the phrase.

The final layer of processing is the **semantic** layer (figure 15). The main task of this layer is to fit the jigsaw pieces together into larger coherent structures which can be passed to the data mining. It achieves this by using semantic subcategorisation – key relational words (such as *grant*) know the semantic types of the arguments they expect, and the semantic layer scans the sentence searching for likely candidates. It uses a combination of quite simple 'eager' rules, recursively applied, and exceptional patterns to cope with more difficult cases.

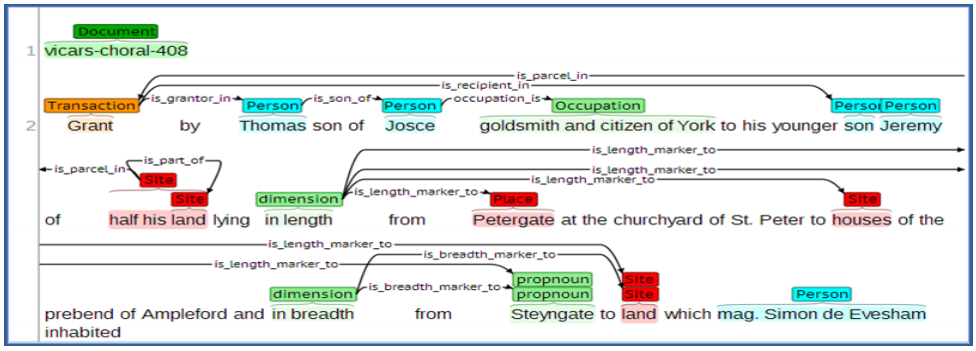


Figure 14: The semantic layer: build semantic relationships by gluing the pieces together.

E) EVALUATION

A comparison between the manual annotation for this sentence and the automatic output can be seen in figures 16 and 17. Many of the key relationships have been successfully identified by the NLP system, but there are also some interesting mistakes. The NLP system has concluded that Josce is a goldsmith, rather than Thomas – this is in fact a more natural interpretation in modern English, but this charter comes from a time where surnames were beginning to evolve, so that *"Thomas son of Josce"* would have been seen as close to a fixed phrase. It also has not quite worked out that Jeremy is a son of Thomas, although it knows that the son was the recipient of the land. There are also one or two 'internal' types, such as dimension and propnoun, which ought to be private to the NLP system but have appeared in the final output (not incorrect, but not recognised as part of the official semantic language).

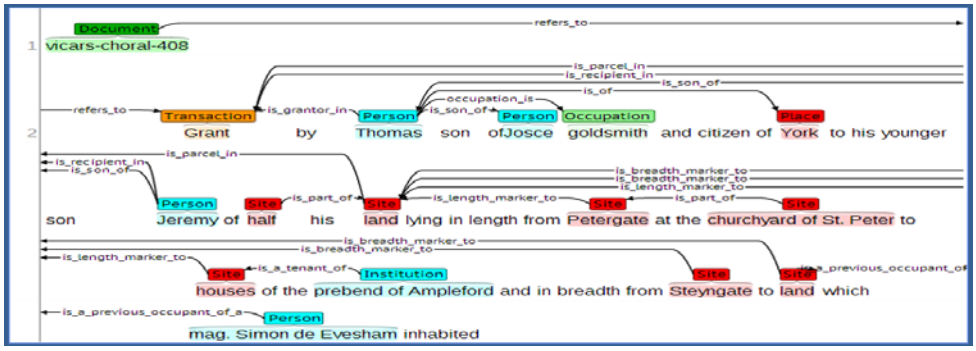


Figure 15: The manually created annotations

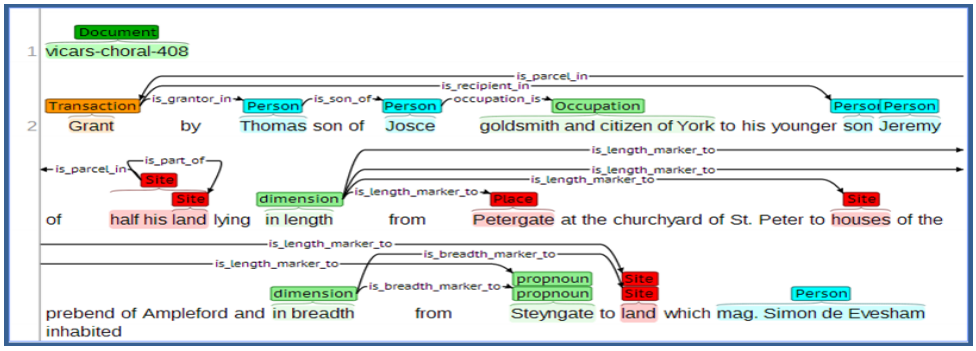


Figure 16: The automatically generated annotations.

Development of systems at this level can always be subject to incremental improvements. At the time of writing, a final round of adjustments to resolve some of these issues is still in progress, which will be followed by a proper formal evaluation of the NLP system performance, to be included in the final version of the White Paper.

We will also include some examples of Latin processing, which have been somewhat delayed by personnel changes and illness, and so is slightly overrunning at this time.

5. SOFTWARE

The NLP software for ChartEx exists as a freestanding system based on SWI Prolog and Java that runs on Windows and Linux. As soon as we stabilise the final project version, we will make it more widely available under a suitable open source licence (possibly Apache 2, is that is compatible with the OpenNLP licence). The codebase will continue to be developed under further projects, including an AHRC 'Big Data' project which will allow our collaboration with Leiden to continue, so we are optimistic of being able to provide some degree of ongoing support and development for the use of this software.

6. RESEARCH QUESTIONS: HOW HAVE WE DONE?

1. **To develop and deploy a system that combines NLP and DM to extract data useful to researchers regarding locations and related actors, events and dates from digital charters**. **We have developed a system which extracts useful and usable data. We have not been able to fully deploy that system, although the key components have been tested in various combinations covering all the functionality required for deployment.**
2. **(not applicable to NLP).**

   3. To investigate whether the ChartEx system can produce efficient and accurate knowledge from charter documents by processing the information in English summaries of the charters in comparison to the full Latin text of the charters. We know that some of the English summaries were abridged in ways that removed essential information for ChartEx. For the project, TNA recovered some of this lost

information for us, but if this problem is more widespread, it seems likely that the Latin would be a more reliable source, if it can be processed sufficiently accurately. But we have ot carried out systematic tests yet due to delays with the Latin component.

4. **To investigate whether the rules used with Latin charters of UK provenance can be applied to Latin charters from different parts of Europe, or if not, how much modification of the rules is needed in order to produce accurate relationships. Not done yet, but should not be an enormous task once the system is ready. (ie not given up on it).**

7. REFERENCES

Evans, R. (2013) The Extended Lexicon: language processing as lexical description In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, Hissar, Bulgaria, 7-13 September, 2013.

Evans, R. and Gazdar, G. (1996) “DATR: a Language for Lexical Knowledge Representation.” *Computational Linguistics* , 22(2), pp. 167-216.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of EURALEX 2004*, Lorient, France, pp. 105-116, (http://www.sketchengine.co.uk)

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J. (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. (http://brat.nlplab.org/)

C. DATA MINING: RECONSTRUCTING MEDIEVAL SOCIAL NETWORKS FROM ENGLISH AND LATIN CHARTERS

(Dr Arno Knobbe, Marvin Meeng, LIACS, University of Leiden)

1. CHARTERS

As described above the charters in our collection record transactions of property. Typically, they contain fairly concise descriptions of the grantor and recipient of the transaction, some description of the property involved, often made more precise by mentioning the previous owners, and finally a list of witnesses. The following is an example of a charter from the Vicars Choral collection (pertaining to properties in the York, UK area):

> 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]

Note that this charter primarily identifies two people, Thomas, son of Josce, goldsmith and citizen of York and Jeremy, his younger son. The other person mentioned in the body text is mag. Simon de Evesham (mag. for magister, an academic title at the time) does not play a direct role, but is mentioned to specify a piece of land that is needed to identify the specific property being transferred here. The transaction relates to half of the land previously belonging to Thomas the goldsmith, and is mostly identified by how it is positioned in relation to other properties or landmarks, that were perhaps easier to identify at the time. Note that Petergate and Steynate are crossing streets that still exist in York as Petergate and Stonegate. The list of witnesses offers some clue as to the people involved, but does not play a major role in our work. Finally, the document is dated fairly accurately, but this is definitely not always the case, and differs from collection to collection, becoming common only after c. 1300.

One important thing to note here, that plays an important role in the process of record linkage, is the fact that most information conveyed in a charter is in natural language, something which hinders direct interpretation of the documents and leaves room for ambiguity. For example, one could argue in this text that Josce is in fact the goldsmith and citizen of York, rather than Thomas. Additionally, there is no notion of registered land or geographic coordinates, nor do people have social security numbers, as one would expect in modern legal transactions. To make matters worse, there was no unified spelling of people and place names, such that a considerable level of flexibility will have to be assumed when matching names across documents. Also, the notion of last names was only slowly appearing in Medieval England, such that people often are only identified by their first name. In many cases, people's origin or profession served as last name, such as with William de Gerfordby or Robert le Spicer, but these 'last names' did not serve as family names. Needless to say, the unequivocal matching of people and sites both within and across charters is a challenge.

2. RECORD LINKAGE

Despite the hurdles mentioned in the previous section, linking people and sites across charters turns out to be quite doable. For one, the population at the time was much smaller, and the people involved in property transactions is only a fraction of that, since not many people could afford to own land. Additionally, the charters tend to provide sufficient 'circumstantial evidence' in order to recognize people, perhaps as an unconscious attempt of the author to be sufficiently specific. For the charter mentioned above, Tomas son of Josce goldsmith and citizen of York is actually a phrase that appears in another charter, as well as the son Jeremy, although it appears both as Jeremy as well as Jeremias there. This charter also mentions that Jeremy is the son of Mariot, who is (by then) the widow of Thomas. This demonstrates how a social network slowly appears when being able to link individuals across charters. The links in this network can represent family relations explicitly mentioned in the document, but also the property transactions themselves: a charter connects the grantor and recipient. In the networks presented, we also include might-be connections, such that we can communicate persons mentioned in multiple charters, with various degrees of certainty.

Our record linkage method combines a probabilistic approach with a certain level of logical reasoning. The probabilistic side of our method aims to determine whether a candidate link (two mentions in two charters refer to the same person) is very probable, given the evidence available in both charters. Generally, for each person (a subset of) the following information is available: first name, occupation, title, last name, family relations. For each matching item (ignoring the complications of spelling variation for now) between the two persons, we need to determine the probability of making a wrong assumption of identity, and combine these probabilities in an overall confidence score of the assertion that we are dealing with one and the same person. Obviously, the more pieces of evidence we have and the more reliable that evidence is, the higher our estimated confidence.

One of the big challenges here is to estimate probabilities for individual items. For example, finding a Thomas in two charters may not be much evidence, if Thomas is a very common name at the time. Therefore, we need to estimate the frequencies of all first names in order to compute the partial probability. We opted to do this in ChartEx simply by using the combined collections as a source of names statistics, producing a histogram of all names appear in the collections. The five most common names found here, in descending order, are John, Thomas (unfortunately), Robert, William, Richard. Josce appears only twice, making the match in the previously-mentioned charters much more probable than the one concerning Thomas. The same process can be repeated for the occupations (yeoman, gentleman, esquire, clerk, goldsmith, ...). For family relations, one can simply adopt the same reasoning as for first names. Knowing one's father is called Josce is just as informative as being called Josce oneself, so we can simply use the first name statistics. The process of determining the probabilistic contribution of a last name was less clear, for reasons mentioned earlier. For the lack of reliable statistics on last name occurrences, we simply introduced a fixed score, such that a matching last name contributes to the confidence by a constant amount.

Aside from the probabilistic reasoning described here, there was also a considerable amount of logical reasoning, notably when conflicting evidence was present. For example, having a mother with a different name is problematic, regardless of the other matching evidence. The same is true for appearing in two charters that are separated by more than a

hundred years, although this reasoning is less clear-cut when the separation reduces to say 30 years. For lack of a good model of longevity in the Middle Ages, and more importantly in what age bracket one might be expected to be involved in property transactions, we introduced a fairly simplistic probability function depending on the number of years between the charters, with all separations over 80 years being discarded. Some more sophisticated reasoning, that reflects some of the logic of charters and of the era, was involved when considering when considering relations that potentially do not last, or change over time. For example being married to person A doesn't exclude one from being married to another person B at a later stage, or being one's widow. As an aside, expressions such as son of and mother of were used to infer some of the genders of lesser-known names such as Fange (male) and Thomasina (female).

Although the probabilistic approach taken produces very satisfactory results, replacing considerable manual labour by historians, it has a few drawbacks that it shares with many probabilistic approaches to record linkage. First of all, producing name and occupation statistics from the collections themselves introduces a certain bias, for the simple reasons that people may appear more than once. Especially the names of those who own a lot of property (e.g. Simon de Evesham) will appear higher in the ranking than is realistic, with the undesirable side effect that their matching actually becomes less confident. Sometimes, assumptions need to be made that are not supported by sufficient data, such as in the case of last names. Finally, a common complaint of probabilistic approaches is that the combined estimate of confidence assumes that the individual probabilities are independent, which they are often not. For example, first and last name frequencies are known to be quite dependent, although this example doesn't apply to our data. It does however apply to first names and occupations, which are not independently distributed. Despite some of these drawbacks, the method appears to work sufficiently well, and as long as one doesn't interpret confidences as absolute numbers, but rather as rankings, the confidence numbers are very usable.

3. QUANTITIVE DETAILS

In total, five collections of charters were available to the project, being:

- The Vicars Choral (University of York), 125 charters mannually annotated, English, 5,000 charters (dated).

- Borthwick (Borthwick Institute, University of York), 55 charters manually annotated, English.

- DEEDS (University of Toronto), 49 charters manually annotated, Latin, over 10,000 charters.

- Wards2 (The National Archives, UK), 48 charters manually annotated, English, 7,000 charters.

- Cluny (University of Columbia), 50 charters manually annotated, Latin, over 5,000 charters (dated).

In the English documents, a total of 112 different first names occur, where we assume different spellings are different names. Of these names, the gender of over 85% could be inferred from the context in which they appeared (for example Thomas, son of Josce implies Thomas is a male name). Of the names for which the gender was resolved, 36% was female.

It should be noted though that in absolute sense, women were much less mentioned than men, especially where ownership of property is concerned. In a ranking of names according to their frequency, the first female name (Margaret) appears at rank 15. Also, the common name John is over 17 times more common than Margaret. This Medieval gender difference is also indicated by the occupation statistics, where the first clear female 'occupation' (an annotation that was used somewhat liberally in this project) is 'widow' at rank 12, after clearly male occupations such as 'yeoman' and 'esquire'.

4. OUTLOOK

The ChartEx project has by now finished. The funding for ChartEx was relatively short, making all steps in the process somewhat proof-of-concept. Still, a working system was produced that allows historians to work with large collections of charters in an integrated manner. The record linkage activities continue in a new, somewhat larger project with The National Archives (UK), where not only medieval records are involved but also more modern civil records. There are also well-developed ideas for a follow-up project to ChartEx, called The Medieval Mine, which aims to exploit the new capabilities of analysing collections in their entirety, and mining the structured result by means of modern Data Mining techniques.

Figure 17: Charter 408 from The Vicars Choral, in relational form. Note that this network does not yet involve record linkage across charters.

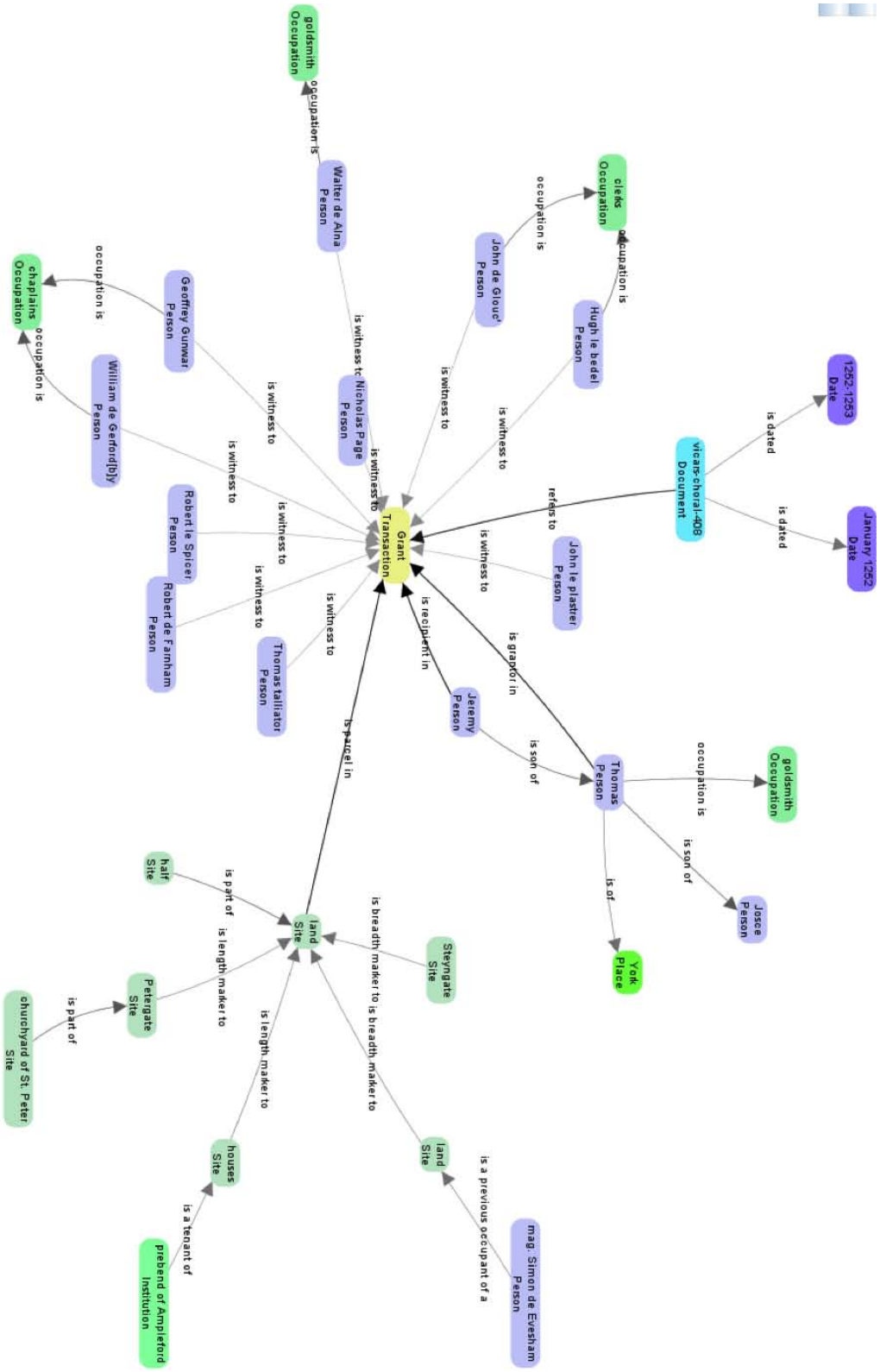
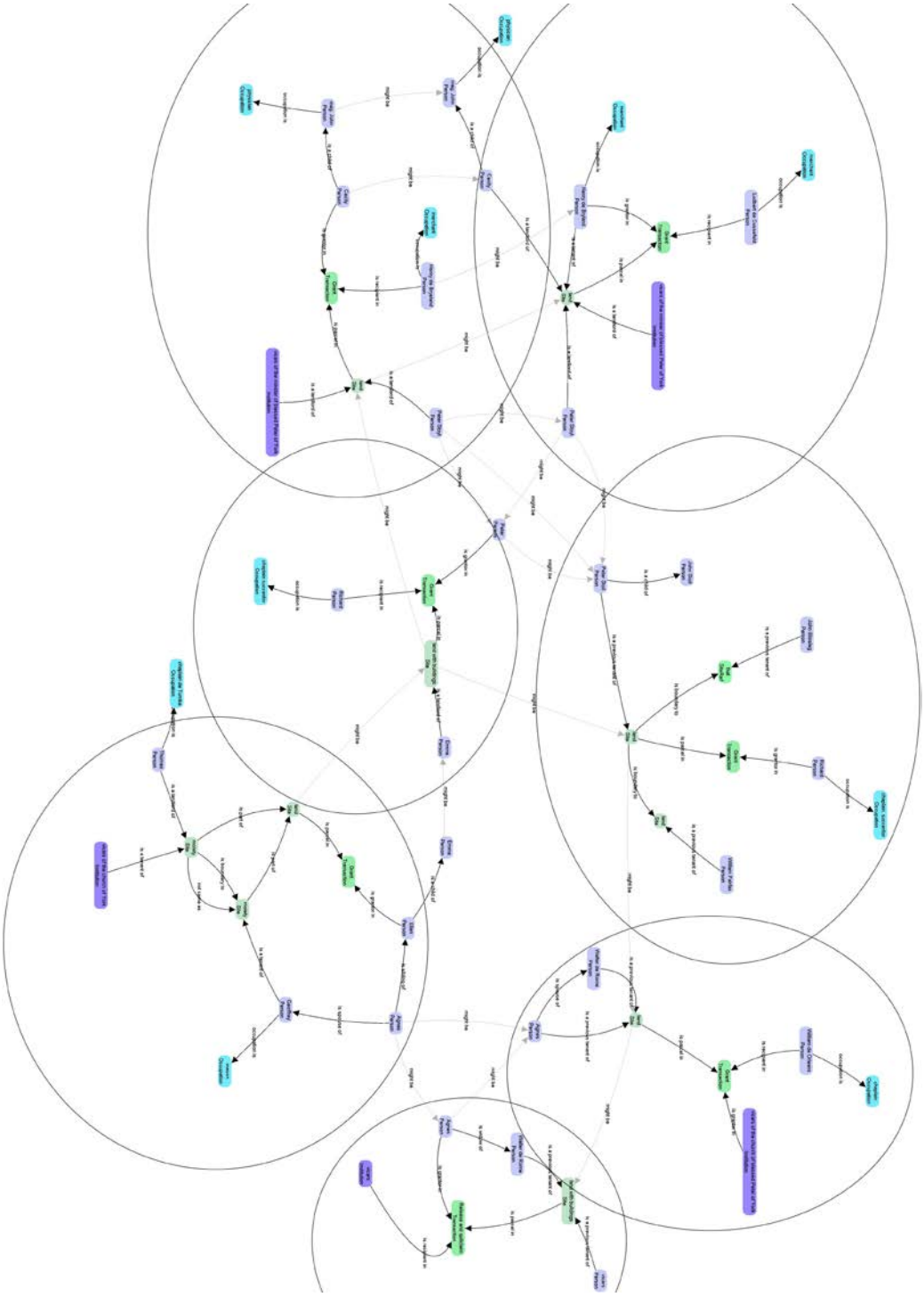


Figure 18: Partial network generated from a subset of charters. The ovals roughly cover the seven charters involved. Gray lines indicate hypothesised links between people in various roles in the charters. Note some of the spelling variations.



D. THE CHARTEX VIRTUAL WORKBENCH

1. INTRODUCTION

The overall aim of the ChartEx Project was to develop new ways of exploring the full text content of digital historical records. One of the specific aims outlined originally was:

> *"To investigate whether researchers working with a virtual workbench based on novel instrumental interaction techniques will produce more useful knowledge from charters than a human working alone or an automated NLP/DM system working alone"*

This section documents the progress towards this aim, including the initial requirements gathering process, the design and implementation of the ChartEx Virtual Workbench, and an evaluation of the ChartEx Virtual Workbench with potential users.

2. ESTABLISHING USER REQUIREMENTS

The ChartEx Project has demonstrated how Natural Language Processing (NLP) and Data Mining can be used to automatically extract information about places, people and events from medieval charters and find new relationships between these entities. To allow historians, archivists and other relevant parties to explore and interact with this information, it was necessary to develop a robust, innovative interface that would go far beyond current methods of exploring the data without disrupting existing working practices. This required a thorough understanding of how historians, archivists and other relevant parties currently interact with the data and what their requirements are for tools that would support them in exploring the information.

A) METHOD

This aspect of the investigation began with a series of in-depth contextual inquiries with 7 historians, which allowed the complex work practices of this cohort to be explored. The contextual inquiry methodology was developed by Beyer and Holtzblatt (1997) and incorporates a one-on-one observation of work practice in its naturally occurring context. Detailed information can be collected about work practices by observing and interviewing a participant whilst he or she actually works. At any point before, during, or after an observation, the interviewer can discusses the participant's daily routines and work processes in order to develop a deep understanding of them. Ultimately, the purpose of contextual inquiry is to understand how and why something is done or why something is not done and how it may be improved.

PARTICIPANTS

An initial round of contextual inquiries was conducted with seven participants. 3 were male and 4 were female. All of the participants were historical researchers currently working in universities and institutions in the UK, the USA, and Canada.

PROCEDURE

Participants were interviewed face-to-face in their place of work. They were given an introduction to the ChartEx Project and its aims and assured of the confidentiality of the information they would be providing. Both audio and video were recorded for later transcription. Contextual interviews lasted between 90 and 120 minutes. When the interview schedule was complete, participants were asked whether they had any further comments they would like to make. Finally they were thanked for their time and fully debriefed.

B) RESULTS

The contextual inquiries revealed that historians work in very different ways and consequently have very disparate requirements. For example, some historians choose to search for information and require very precise means of constraining the search space. Others prefer to browse the data and require the ability to link from one historical document to the next. Some historians are more visual, choosing to work directly with maps, visualisations and other abstractions of the data, whereas others are more rooted in the actual documents and text within them. Nevertheless, the investigation identified a number of distinct themes that reflect the current working practices of historians.

SEARCHING

Many of the historians expressed the need for a search facility that would allow them to search the data for common phrases and relevant keywords, similar to an Internet search engine. In addition to such a freeform search, many of the historians also wanted the ability to constrain the search to specific types or sources of data. For example, a number of them said they would like to search for specific locations or sites. Others would prefer to constrain their search to a particular transaction that occurs in the data or the people involved in that transaction. Some historians tend to work within particular collections of charters and so it was important that the search could also be constrained in this way.

- **Requirement:** To be able to search for phrases, keywords and entities across documents
- **Requirement:** To be able to constrain searches to particular entity types or collections

Another important requirement with regards to searching the data was that the system be sufficiently tolerant of different spellings and descriptions. Variations and errors in spelling are common in historical documents, as one of the historians attested: *"You may find the same person on one occasion mentioned as Smith, and in another as Faber – Faber is the Latin word for Smith, this happens a lot. Or he might be Johannes Fuller and in another he might be*

Johannes Fullonis. So the Latin and the English get quite mixed up." One of the historians proposed that the system could also suggest alternative spellings, similar to Google's *"Did you mean?"* facility.

- **Requirement:** The system must be tolerant of different spellings and variations of names as well as errors in entry

DOCUMENT EXPLORATION

Key to the work of many of the historians is the ability to and explore and work with the actual text of documents, as opposed to a summary or abstraction. The descriptive text contained within documents might allow the researcher to identify persons or properties, or the entities involved in a transaction and their relationship to each other. This was summarised by one of the historians, who said: *"I go back to my original transcript all the time. And I think that is part of the historical process. The more you learn and the more you do, when you go back to something, you see something different. So actually yes, having that actual word-for-word transcript is really important."*

- **Requirement:** To be able to view the original document text

Historians work with many documents at a time (tens, hundreds, sometimes thousands of documents) and so the ability to open multiple documents at once and identify connections between them is crucial. One of the historians commented: *"I looked at many documents, and also I looked at documents that were outside the close in Petergate, because they are abutting closes, which tell you where the boundary really is."* Not only should the system allow multiple documents to be opened, it should also allow users to switch between the documents with ease. One of the historians said: *"you kind of need to see two charters together sometimes in order to make sense of things"*.

- **Requirement:** To be able to compare more than one document at the same time
- **Requirement:** To be able to work with large numbers of documents

Historians are interested in different aspects of the information contained within documents. Several of the historians expressed the requirement to be able to highlight and distinguish different types of information within a single document. This might be the people mentioned in a document or the events the document describes and the dates they occur. One historian commented that when reading documents they *"look through that document and extract all the information I possibly can … The idea is that as you get better at these things you can perhaps skim-read them so you don't need to do as much work from them".* Highlighting different categories of information within a document text would allow historians to process the document more efficiently.

- **Requirement:** To be able to highlight different categories of information with documents

VISUALISATION

Many of the historians work closely with existing maps or create their own diagrams and visual representations of the data. For example, one of the historians described how they create *"little two-dimensional images of the logical relationships"* between entities within a set of documents. Another historian placed extracts of texts on a timeline. Another described how they used a pinboard with colour-coded pins to identify professions and locations on a map. To support this activity, many expressed the need for interactive visualisations of the data. One historian felt that being able to view information within a visualisation might reveal patterns that may not be obvious from looking at the individual data: *"I was so excited by that nodes diagram* [an example visualization shown to the participant]*, because it isn't something I can make myself, but it is a way of quickly pulling out and rearranging information I already have, but maybe makes connections more readily … It is much harder to see the connections between this material when it is in this format"*.

- **Requirement:** To be able to visualise information in a node diagram and be able to reorganise the nodes to show different relationships
- **Requirement:** To be able to view different categories of information (e.g. profession) using a colour coding scheme

IDENTIFYING CONNECTIONS

Another key aspect of the work of historians is identifying and exploring connections both within and between documents. Medieval charters record legal transactions of property of all kinds: houses, workshops, fields and meadows and describe the people who lived there. These entities may appear across numerous documents, allowing historians to identify connections and build an understanding of the complex relationships within them. One of the historians explained: *"I have identified properties by looking at the names. And this William Blanchard, you can find out he lived in previous properties, then he moved, and you can follow all these documents looking at the previous leasee, the previous occupier. Names are very important, because they connect property. You can say my neighbour lives there, and my neighbour was this person".* Another historian explained how properties are often described in relation to neighbouring properties or previous occupants. Lease documents typically include references to previous documents about the property, allowing the researcher to link entities over time: *"Sometimes you link property because of the name of people who lived there. And also it tells you a lot about how the neighbourhood changed"*.

- **Requirement:** To be able to draw upon information from across documents and within documents

- **Requirement:** To be able to cross-reference information across different documents
- **Requirement:** To be able to reference one document from another

ADDING METADATA

In addition to exploring documents and identifying connections between then, many of the historians also expressed a desire to contribute to the data by adding and sharing additional metadata. This may be to support and strengthen a particular connection that has been identified between entities or it may be to refute a relationship to avoid further confusion. Many of the historians explained how they would like to add new information to a document: *"Each historian could add information and links between things so that other people coming along later would benefit from that"*. One also felt it was important to be able to highlight *"the bits that you don't know about"* to identify where information is lacking. A number of historians also explained how they use paper, Word documents and Excel spreadsheets to record notes and information for their own reference: *"The status of the analysis is complete but this was for myself, as notes about my work, and nothing to do with the analysis of the document."* They expressed the need for a system that could incorporate this additional information.

- **Requirement:** To be able to supply additional evidence and information
- **Requirement:** To be able to highlight information that is lacking about something
- **Requirement:** To be able to include notes for the user's own reference

The use of contextual inquiries has allowed the complex work practices of historians to be explored. This has resulted in a number of requirements for the ChartEx Virtual Workbench that would allow historians to search for documents, explore documents, view visualisations of the data, identify connections between entities and provide additional metadata.

3. DESIGN AND IMPLEMENTATION OF THE CHARTEX VIRTUAL WORKBENCH

A) INITIAL PROTOTYPE

Following the user requirements process, a lo-fidelity, wireframe prototype was developed incorporating many of the identified features and requirements. This initial prototype was shared with members of the ChartEx Project as part of a participatory design workshop. This is a process that involves developers, potential users and other relevant stakeholders working together to design a solution.

The initial prototype took the form of a series of panels, each with its own set of tabs. In one of the panels, users can search for documents and view the results. Users can also open individual documents within new tabs and view the document text, any associated images, as well as metadata about the document. Users can toggle highlighting of entities within the document text and each highlight links to more details about that entity. In a separate panel, users can view details about individual entities, including their relationships to other documents and entities, and associated confidence ratings. Users can also view visualisations of the relationships between entities and between documents.

The participatory design workshop proved successful in identifying aspects of the design that the historians found particularly useful (or not) and establishing the requirements for further iterations of the design.

B) IMPLEMENTATION

DEVELOPMENT TECHNOLOGIES

The ChartEx Virtual Workbench application was developed using HTML, CSS, JavaScript (including the jQuery and jQuery UI libraries) and PHP within the CodeIgniter framework, running on an Apache server with a MySQL database. The visualisations in the application are based on the radial graph (RGraph) visualisation from the JavaScript InfoVis Toolkit.

FEATURES OF THE CHARTEX WORKBENCH

The application comprises three distinct panels: Search, Documents, and Entities (see Figure 19). The Search panel is where users can search for documents or entities, constrain their searches either by collection or entity type, and view their search results. The Documents panel

is where users can view individual documents, including the document text (with highlighting of entities) and a visualisation of the transactions outlined in the document. The Entities panel allows users to view individual entities, including related documents and entities and a visualisation of the relationships between entities. The various features of the ChartEx Virtual Workbench are described below.

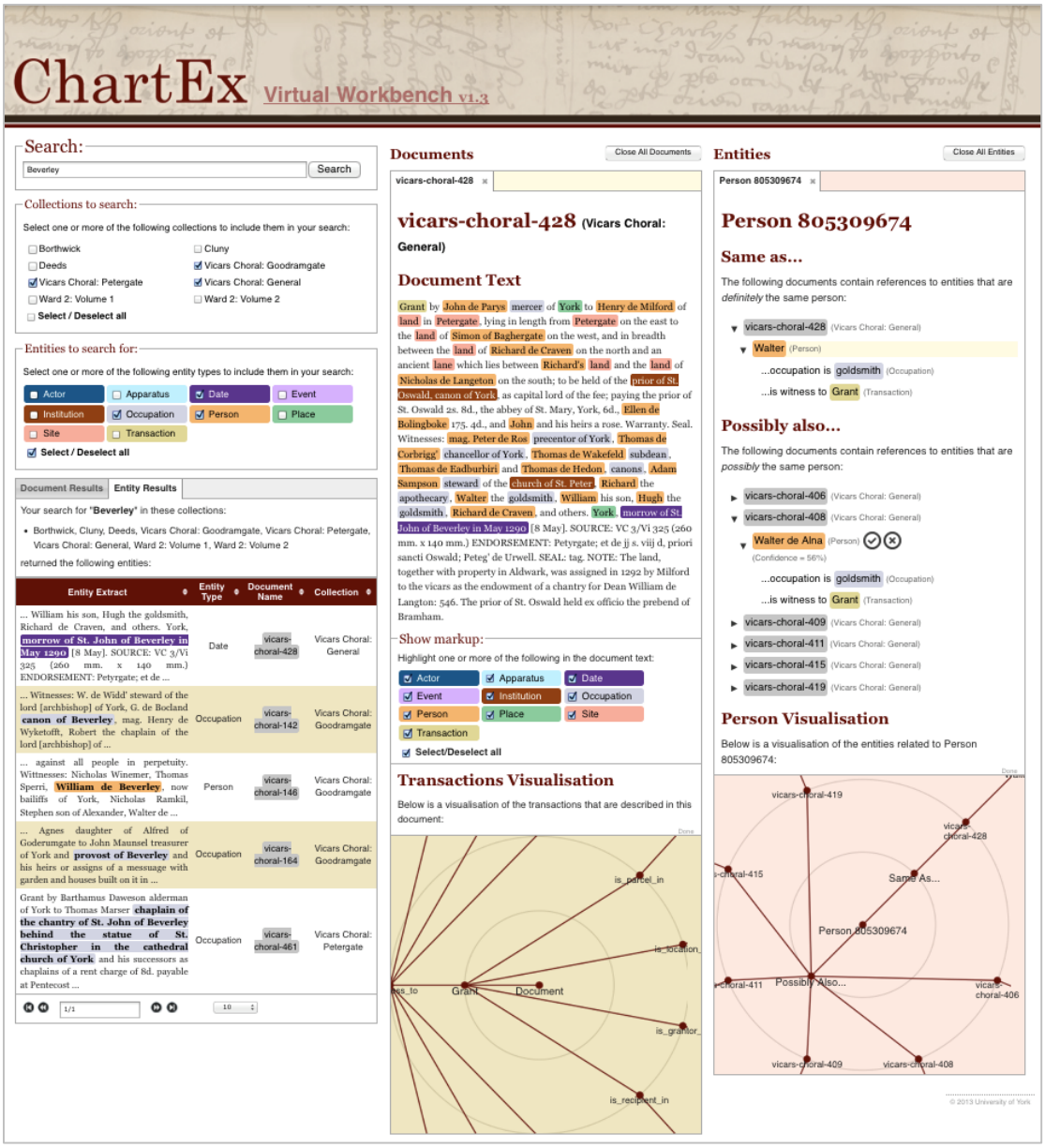


Figure 19: Overview of the ChartEx Virtual Workbench

SEARCH

The "Search" feature comprises a freeform text field and a Search button (see Figure 20). Users can search the data by entering phrases (e.g. *"land to the west of"*), keywords (e.g.

"churchyard") or the names of known entities (e.g. the placename *"Beverley"*). This satisfies the requirement for users to be able to search for phrases, keywords and entities across documents.

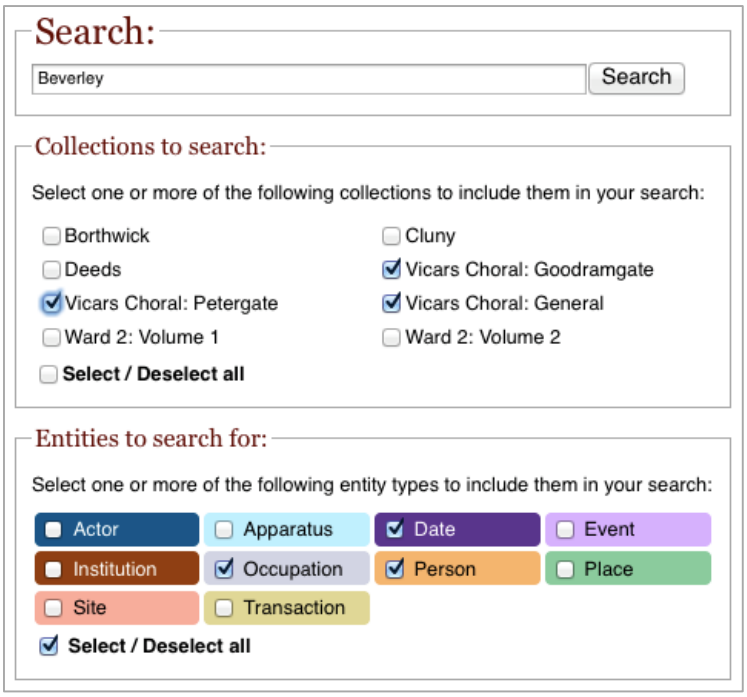


Figure 20: The "Search" feature

Searches can be also constrained by collection and/or entity type (see Figure 21). In the *Collections to search* and *Entities to search for* fields, users are presented with dynamically-generated lists of checkboxes representing the different collections and entity types within the database. Each field also includes an additional checkbox to allow users to select or deselect all of the options. Users can select the collections and entity types they are interested in before running a search to constrain their results. This satisfies the requirement for users to be able to constrain searches to particular entity types or collections. Although the application will not tolerate different spellings and variations of names, it will match partial search strings (e.g. a search for *"Bev"* will match *"Beverley"* as well as *"Bevley"*) allowing some degree of flexibility.

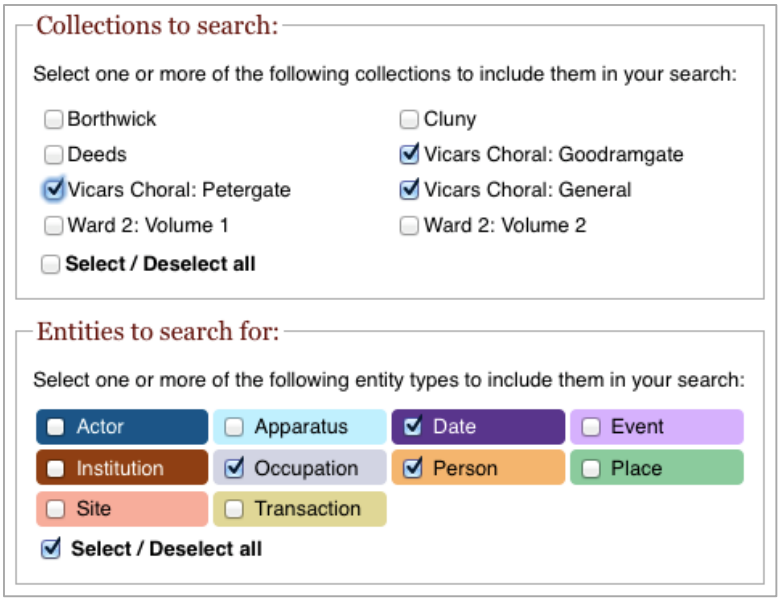


Figure 21: The search for "Beverley" is constrained to dates, persons and occupations in the Vicars Choral collections

Search results are presented immediately below the search options. These are divided into two tabs: “Document Results” and “Entity Results”, reflecting the different ways in which users want to explore the data. The “Document Results” include any documents in which the user’s search query matches part of the document text, regardless of whether it has been identified as an entity by humans or by the NLP algorithms. The “Entity Results” are similar, except they only include documents in which the search query matches known entities that have been marked up in those documents.

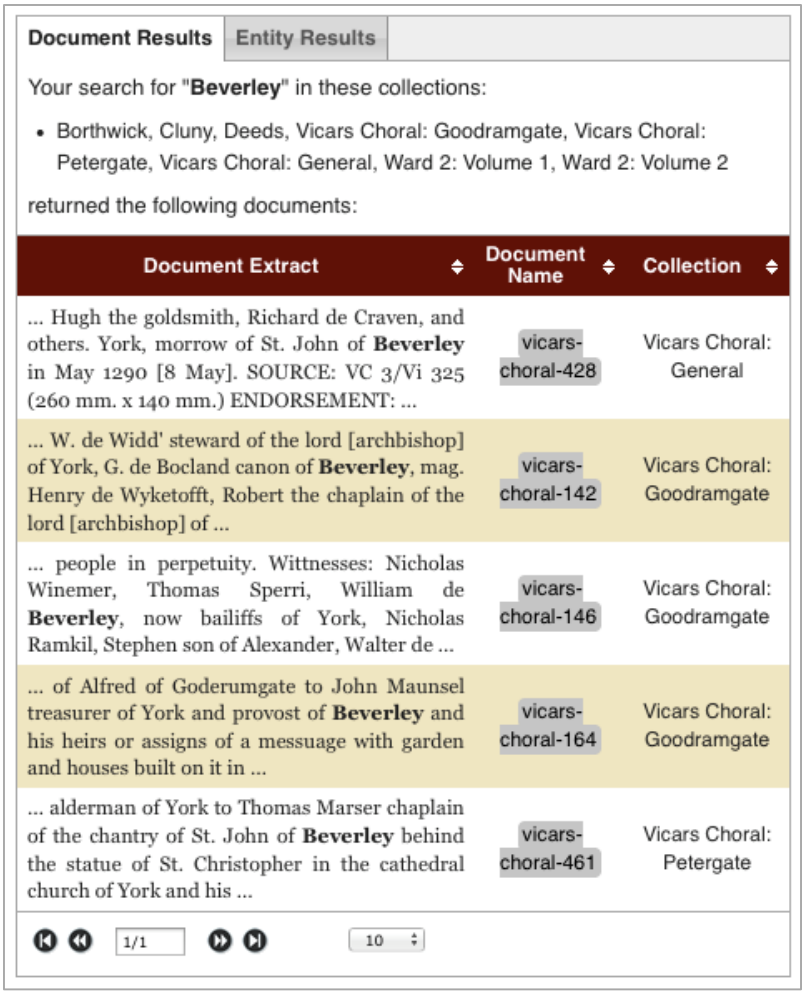
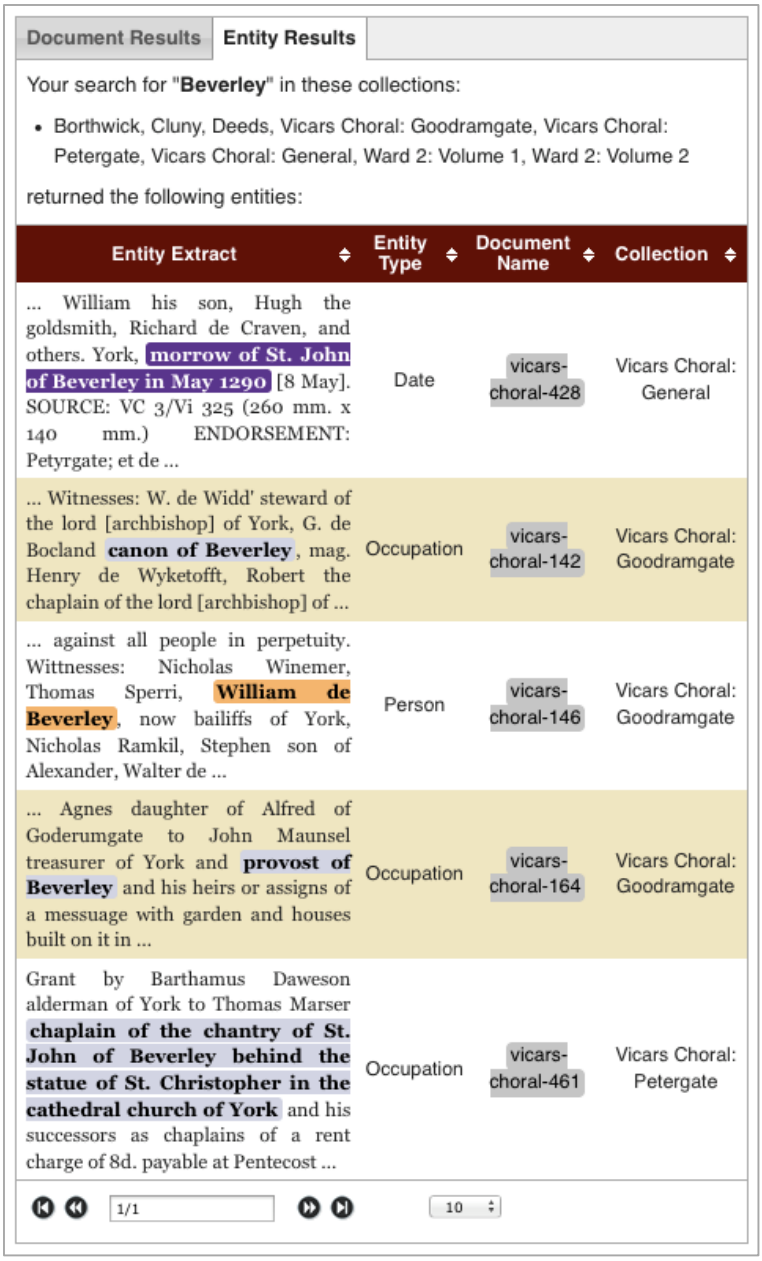


Figure 22: Documents Results, showing the results of the search for "Beverley"

The “Document Results” are displayed in a table that includes: a snippet of the document text that matches the search query (with the search query highlighted), the name of the document in which it appears, and the name of the collection in which it appears (see Figure 22).

The “Entity Results” are displayed in a table that includes: a snippet of the document text that matches the search query (with the entity highlighted), the entity type, the name of the document in which it appears, and the name of the collection in which it appears (see Figure 23). Each table can be sorted by column. To ensure that the results are manageable, searches that return more than 10 results are paginated across several pages. Clicking on a document name will open that document in a new tab in the Documents panel. Similarly, clicking on an entity name will open that entity in a new tab in the Entities panel.



Document Results | Entity Results

Your search for "**Beverley**" in these collections:

- Borthwick, Cluny, Deeds, Vicars Choral: Goodramgate, Vicars Choral: Petergate, Vicars Choral: General, Ward 2: Volume 1, Ward 2: Volume 2

returned the following entities:

Entity Extract	Entity Type	Document Name	Collection
... William his son, Hugh the goldsmith, Richard de Craven, and others. York, **morrow of St. John of Beverley in May 1290** [8 May]. SOURCE: VC 3/Vi 325 (260 mm. x 140 mm.) ENDORSEMENT: Petyrgate; et de ...	Date	vicars-choral-428	Vicars Choral: General
... Witnesses: W. de Widd' steward of the lord [archbishop] of York, G. de Bocland **canon of Beverley**, mag. Henry de Wyketofft, Robert the chaplain of the lord [archbishop] of ...	Occupation	vicars-choral-142	Vicars Choral: Goodramgate
... against all people in perpetuity. Wittnesses: Nicholas Winemer, Thomas Sperri, **William de Beverley**, now bailiffs of York, Nicholas Ramkil, Stephen son of Alexander, Walter de ...	Person	vicars-choral-146	Vicars Choral: Goodramgate
... Agnes daughter of Alfred of Goderumgate to John Maunsel treasurer of York and **provost of Beverley** and his heirs or assigns of a messuage with garden and houses built on it in ...	Occupation	vicars-choral-164	Vicars Choral: Goodramgate
Grant by Barthamus Daweson alderman of York to Thomas Marser **chaplain of the chantry of St. John of Beverley behind the statue of St. Christopher in the cathedral church of York** and his successors as chaplains of a rent charge of 8d. payable at Pentecost ...	Occupation	vicars-choral-461	Vicars Choral: Petergate

1/1 | 10

Figure 23: Entity Results, showing the results of the search for "Beverley"

DOCUMENTS

The Documents panel is where individual documents are displayed. Each document tab includes the name of the document and the collection it appears in, the full document text (with highlighting), controls to toggle document highlighting, and an interactive visualisation of the transaction outlined in the document. Each document is opened in a separate tab allowing users to switch between documents. This satisfies the requirement for users to be able to compare more than one document at the same time. Also, many documents can be opened at once, satisfying the requirement for users to be able to work with large numbers of documents. Users can also close individual documents or close all documents at once.

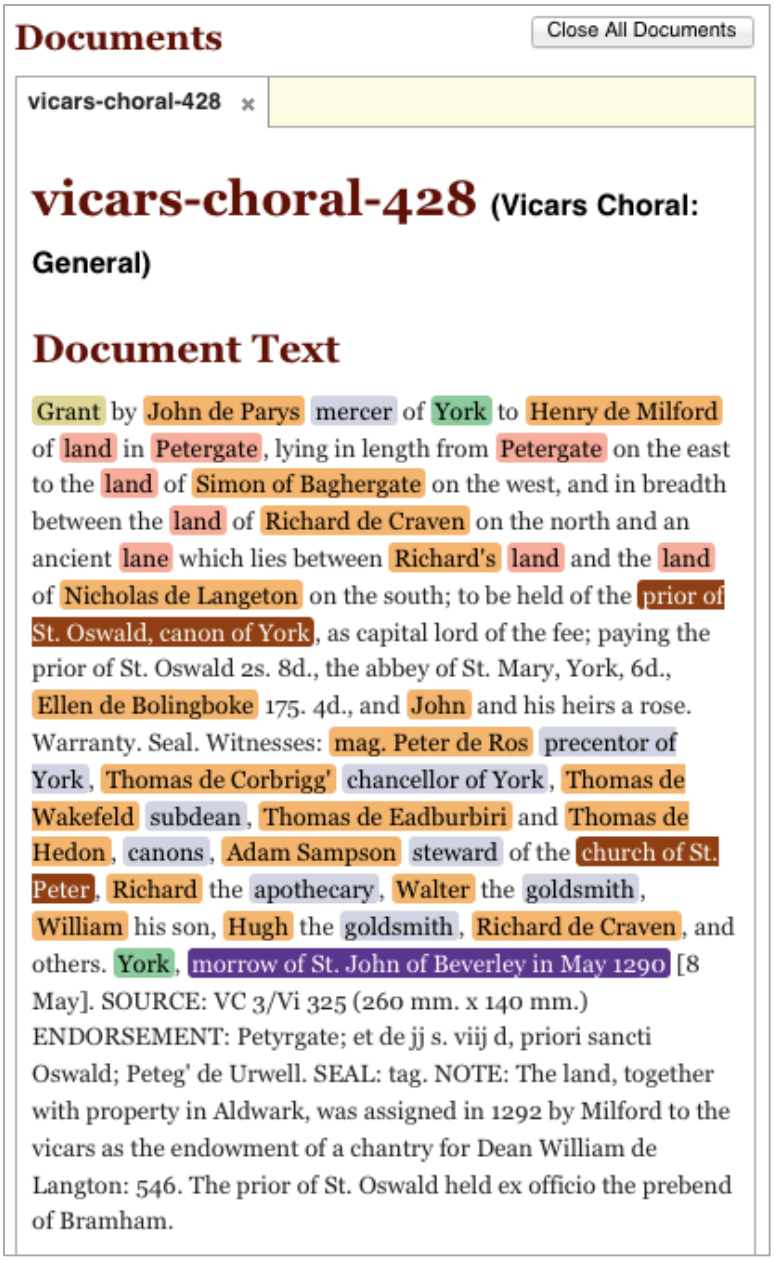


Figure 24: A document tab showing the Document Text

Each document tab includes the full document text (see Figure 24). This satisfies the requirement for users to be able to view the original document text. In addition to this, known entities have been marked up within the document text. Each entity type is colour-coded and users can toggle highlighting of different entity types using the “Show markup” controls (see Figure 25). Here, users are presented with a dynamically-generated list of checkboxes representing the different entity types within the document. Each field also includes an additional checkbox to allow users to select or deselect all of the options. This satisfies the requirement for users to be able to highlight different categories of information within documents. Each entity in the document is also a clickable link that will open that entity in a

new tab in the Entities panel. This satisfies the requirement for users to be able to reference one document from another.

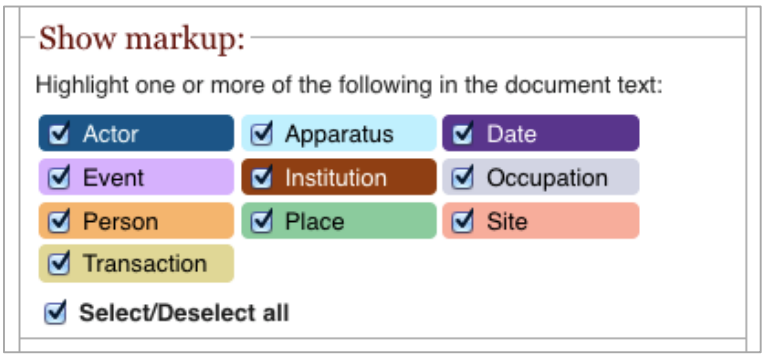


Figure 25: The Show Markup controls used to toggle document highlighting

The "Transaction Visualisation" is presented immediately below the Document Text and "Show markup" controls (see Figure 26). This displays interconnected nodes arranged in concentric circles. At the very centre of the visualisation is a node representing the current document. This is connected to one or more nodes representing the transactions outlined in the document. Each transaction node is connected to another set of nodes representing different relationships in the document. Finally each relationship node is connected to one or more entities that feature in the document. Clicking on any node will centre the visualisation on that node and reorganise the other nodes around it. This satisfies the requirement for users to be able to visualise information in a node diagram and be able to reorganise the nodes to show different relationships. An initial version of the graph was produced with colour-coded nodes representing the different entity types but this proved to be very confusing and distracting.

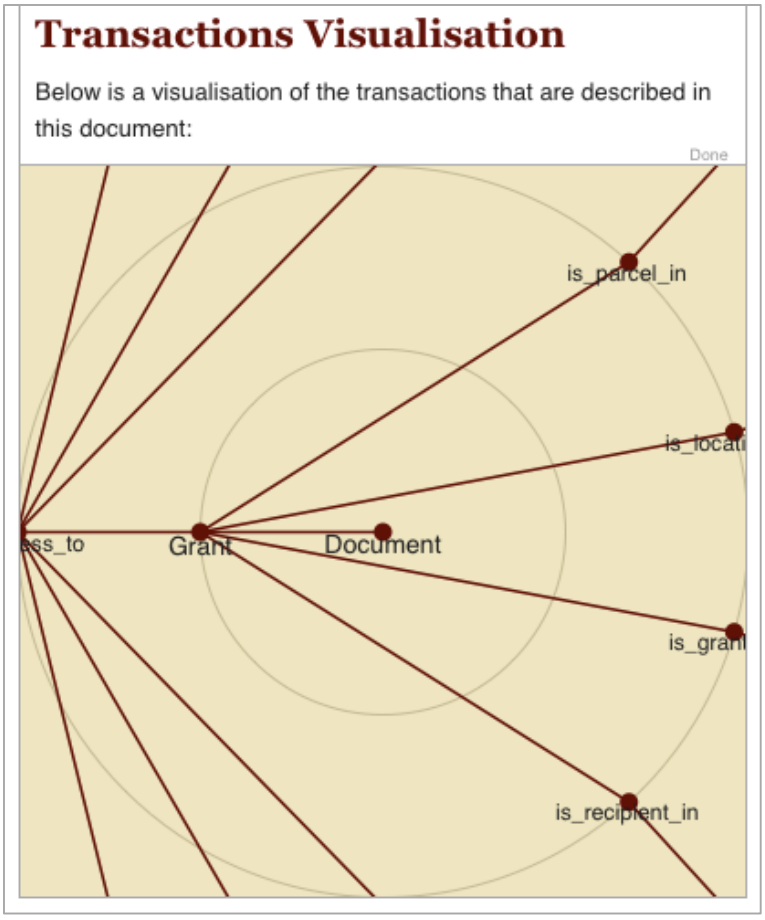


Figure 26: A Transactions Visualisation displaying the transactions outlined in the document

ENTITIES

The Entities panel is where individual entities are displayed. Each entity tab includes the name of the entity, comprising its entity type and a unique id (e.g. “Person 82517” or “Site 26”), a list of documents containing references that are definitely the same entity, a list of documents containing references that are possibly also the same entity, and an interactive visualisation of same as/possibly also relationships. Each entity is opened in a separate tab allowing users to switch between entities and many documents can be opened at once. Users can also close individual entities or close all entities at once.

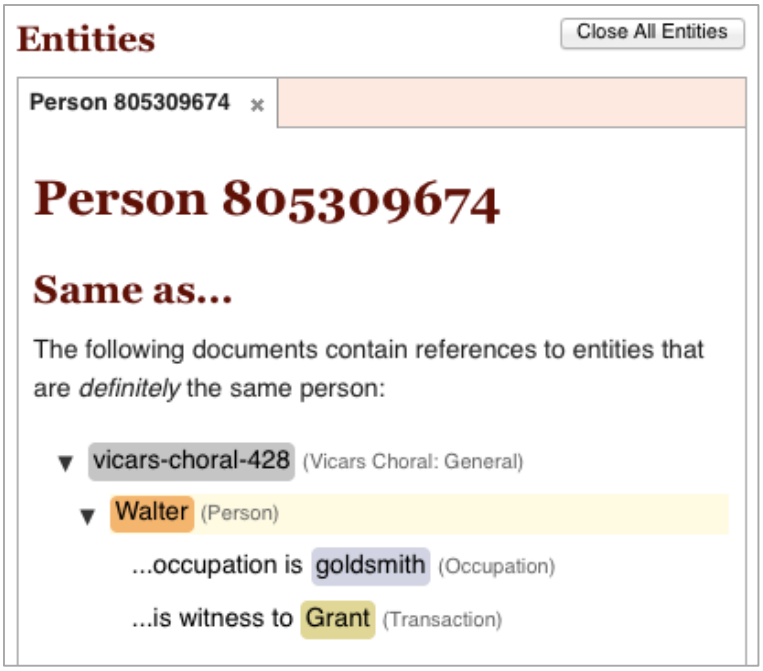


Figure 27: An entity tab showing the "Same As..." relationships

Each entity tab includes a list of documents containing references to entities that are definitely the same entity (see Figure 27). This means that a historian has confirmed these relationships are definitely the same. For example, the person “Walter” definitely appears in the document called vicars-choral-428. As well as the document that the entity appears in, the list also includes the entity name (which may also appear in the document as “he” or “the owner” or some other variation) and other information about that entity (e.g. the grantor in feoffment” or “is of York”). The information is arranged in an expandable tree diagram. This satisfies the requirements for users to be able to draw upon information from across documents and within documents and for users to be able to cross-reference information across different documents.

Each entity tab also includes a list of documents that contain references to entities that are possibly the same entity (see Figure 28: An entity tab showing the "Same As..." relationships). This means that the datamining engine has generated a relationship but a historian has not yet confirmed it. This is displayed in exactly the same way as the “Same As…” list, with the addition of confidence ratings for each relationship to reflect the uncertainty. In addition to the confidence rating, each entry in the “Same As…” list includes buttons that allow the user to confirm or deny that the entity is the same. The “confirm” button causes the entry to bolded, whereas the “deny” button crosses out the entry. This satisfies the requirement for users to be able to supply additional evidence and information. Though users are currently unable to highlight information that is lacking about something, the ability to “deny” relationships between entities offers an opportunity to correct the data. Each document and entity in the

entity tab is also a clickable link that will open either another entity in Entities panel or a document in the Documents panel. This satisfies the requirement for users to be able to reference one document from another.

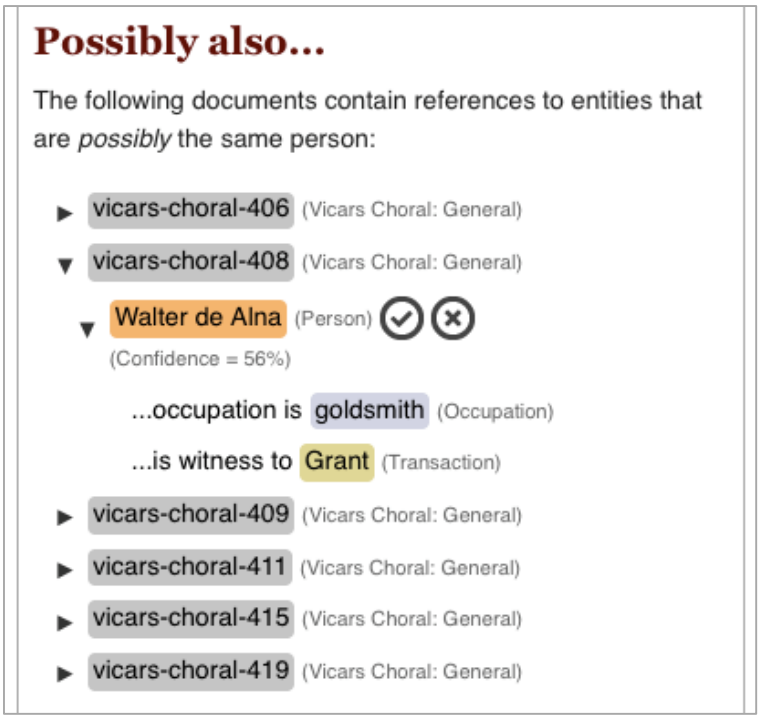


Figure 28: An entity tab showing the "Same As..." relationships

The “Person Visualisation” is presented immediately below the “Same As…” and “Same As…” sections (see Figure 29). As with the Transactions Visualisation, this also displays interconnected nodes arranged in concentric circles. At the very centre of the visualisation is a node representing the current entity. This is connected to a “Same As…” node and a “Same As…” node. Each of these nodes is connected to another set of nodes representing different documents. Finally each document node is connected to one or more entities that feature in the document. Clicking on any node will centre the visualisation on that node and reorganise the other nodes around it. As before, this satisfies the requirement for users to be able to visualise information in a node diagram and be able to reorganise the nodes to show different relationships.

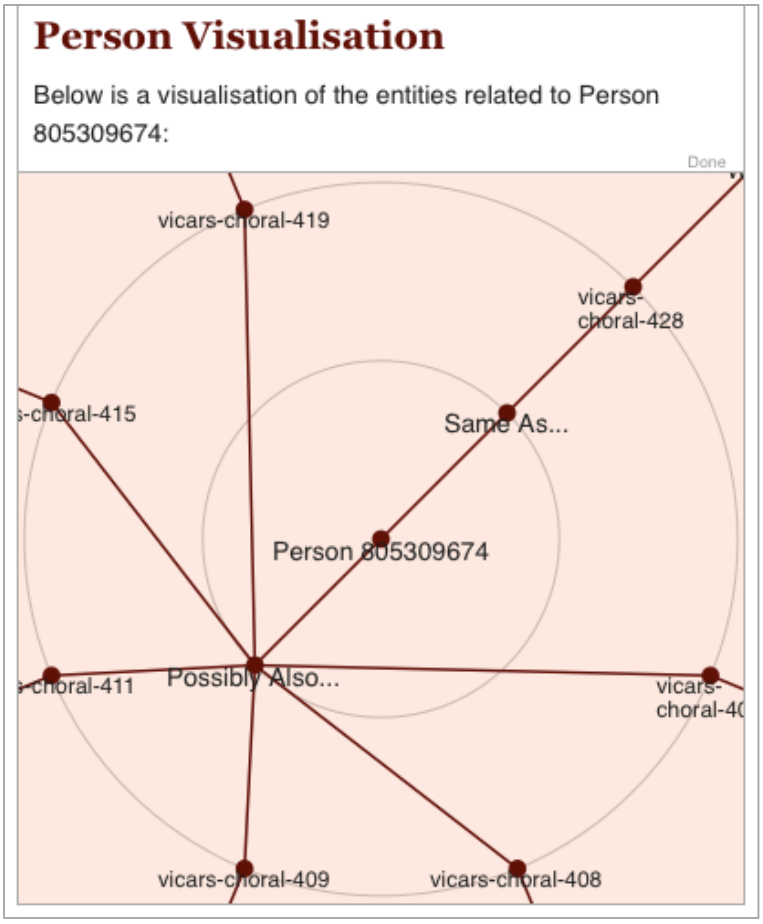


Figure 29: A transaction visualisation showing the current entity and relationships

4. EVALUATION OF THE CHARTEX VIRTUAL WORKBENCH WITH POTENTIAL USERS

This section presents an evaluation of the ChartEx Virtual Workbench with potential users. The evaluation was conducted to determine whether the ChartEx Virtual Workbench is useful and easy to use - primarily by historians but also by other potential user groups, and how well it supports the exploration of medieval charters. The evaluation was also designed to gather important feedback on different aspects of the ChartEx Virtual Workbench, which will be used to further develop the application in future.

A) METHOD

PARTICIPANTS

The thirteen participants for this evaluation were recruited from the staff and student population at Columbia University in New York, USA and the University of York, UK. The participants' mean age was 34 (SD = 10.61, range = 24 - 55 years). 7 participants were male, 5 were female. Nine of the participants were post-doctoral students studying history, medieval history or historical musicology. One participant was a lecturer in medieval history, one was a librarian, and one was a Digital Information Manager and part-time student. All of the participants had completed a Bachelor's degree or higher.

EQUIPMENT

Participants accessed the ChartEx Virtual Workbench using a web browser of their choice. All of the participants used an Apple computer running OS X. The participants completed both the task questions and the follow-up questionnaires on paper.

PROCEDURE

The evaluation was conducted in small groups of 3-4 participants (13 participants in total) in computer laboratories at the respective institutions. Each evaluation session lasted between 1 hour 15 minutes and 1 hour 45 minutes.

After providing their informed consent, the participants were given an introduction to the ChartEx Virtual Workbench. This introduction highlighted the various features of the application and walked the participants through a typical task. The participants were then given a set of 6 practice tasks (which included the answers) to allow them to get used to the application.

Once the participants were comfortable with the application, they were asked to complete a further set of 5 tasks (without any answers provided). These were information retrieval tasks designed to encourage the participants to use different features of the ChartEx Virtual Workbench (although no particular method was enforced). Participants were permitted to ask questions during these tasks but only minimal answers were provided.

As well as space to provide the solution, each task included a set of questions for the participant to answer regarding the ease of completing the task (on a scale of 1-5, where 1 = very difficult and 5 = very easy), their confidence in their responses (on a scale of 1-5, where 1 = not at all confident and 5 = very confident), and how they actually completed the task (open-ended, in their own words). The participants were also asked to rate the usefulness (on a scale of 1-5, where 1 = not at all useful and 5 = very useful) and ease of use (on a scale of 1-5, where 1 = very difficult and 5 = very easy) of specific features of the ChartEx Virtual Workbench as well as provide comments. If participants had not used a specific feature, they were asked whether or not they were aware of the feature. To avoid participants being guided towards the correct or optimal approach to each task, the questions about specific features were printed on a different piece of paper to the task question.

Once they had finished all of the tasks, the historians were asked to complete a demographic questionnaire and a questionnaire about their overall experience of using the ChartEx Virtual Workbench. Following this, they were thanked and fully debriefed.

5. RESULTS

A) TASK PERFORMANCE

TASK ONE

Task 1 required participants to identify the documents within the Cluny collection in which a person called “Aalbert” appeared. All participants found this task to be “easy” or “very easy”, giving it a mean rating of 4.9 (SD = 0.29, range 4-5). Similarly, the participants’ confidence in their responses was very high, with a mean rating of 4.5 (SD = 0.67, range 3-5).

Each of the participants used the search field to search for “Aalbert”, specifying the “Cluny” collection in the *“Collections to search”* field. The majority of participants also specified the “Person” entity type in the *“Entities to search for”* field. The participant who did not use the *“Entities to search for”* field to complete the task was still aware of it. All but one of the participants used the “*Document Extract”* feature in the Document Results to view a snippet of

the text of each document. Similarly, all but two of the participants used the *"Entity Extract"* feature in the Entity Results to view a snippet of the entity in context. The participants who did not use these features to complete the task were still aware of them.

Participants found the "Collections to search" feature very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 4.8, SD = 0.39, range 4-5). One participant commented that more information was needed about each collection. Another commented that the "Select All / Deselect All" option was appreciated.

Participants also found the "Entities to search for" feature very easy to use (mean rating 4.6, SD 0.67, range 3-5) and very useful (mean rating 4.7, SD 0.47, range 4-5). However, one participant commented that they were unsure of the difference between some of the entity types (e.g. "Institution" and "Site") and four participants wondered what some of the entity types represented (e.g. "Apparatus" or "Actor"). One participant said: *"It's not clear to me what selecting them or not contributes to the search. It seems like it is just a colour coding system"*.

Participants found the "Document Extract" feature very easy to use (mean rating 4.7, SD 0.47, range 4-5) and very useful (mean rating 4.8, SD 0.60, range 3-5). One participant felt the feature: *"gives a very good amount of context for the keyword"* whereas another participant felt *"it should be a bit longer"*. One participant said: *"it would be good to be able to select how long you want the extract to be, so it could show more context if needed"*. Another felt that: *"seeing alternative spellings in context is useful"*. One participant commented that is was *"not immediately obvious that you click on document name"*.

Participants found the "Entity Extract" feature very easy to use (mean rating 4.8, SD = 0.63, range 3-5) and useful (mean rating 4.2, SD = 1.03, range 3-5). A number of participants were unsure of the difference between the "Document Extract" and "Entity Extract" features. For example, one said: *"In this case, the results are virtually the same as in "Document Extract", with exception of highlighting"*. One participant said: *"this could potentially be an enormous asset of the program, but it's hard to try out without looking for something specific"* (referring to the relative simplicity of the task).

TASK TWO

Task 2 required participants to identify the name of a person who is always involved in the purchasing of land with "Aalbert" (the person from the previous task). The majority of participants found this task to be "easy" giving it a mean rating of 4.3 (SD 0.67, range 3-5).

Similarly, the participants' confidence in their responses was very high with a mean rating of 4.6 (SD 0.82, range 3-5).

The participants varied in their approach to completing this task. After searching for "Aalbert" once again, seven participants opened each of the documents in which he appears. Five of them then used the "Document text highlighting" and "Show markup" features to look up their response whereas three of them used the "Transactions Visualisation" feature. Three participants obtained their answer from the "Document Extract" feature in the search results. One of these was unaware of another way of completing the task. One participant did not complete this task at all due to a computer problem.

Participants found the "Document text highlighting" feature very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 4.9, SD 0.33, range 4-5). Two of the participants commented that the "Document Extract" feature provided enough information without having to open each document. Another participant would have preferred the text highlighting to be disabled by default.

Participants found the "Show markup" feature easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 4.6, SD 0.79, range 3-5). One participant said: *"It is annoying that "Show Markup" defaults to "Select All" with each document opened. If I only want to display Person and Transaction in each document, I shouldn't have to deselect all then reselect those two"*. Another participant commented that the highlighting feature was *"essential for this kind of task"*.

Participants did not find the "Transactions Visualisation" very easy to use (mean rating 3.3, SD 1.21, range 1-4) and did not find it very useful (mean rating 3.7, SD 1.51, range 1-5). One participant felt it was *"not really useful"* for only one document. Another felt that: *"a bigger window would help see more of the transaction at a glance".* Similarly, another commented that they *"would have liked a full-screen view of this"*. Another two participants were unclear what the concentric circles of the visualisation represented, with one commenting: *"I like the idea of presenting the relations in a kind of map, but I'm not familiar with this concentric format. I'd be more interested in what the relations among the names were - Are any witnesses related? Tenants of the actor etc."* The other said the visualisation: *"is a little confusing - I'm sure once I knew what each ring represented, it would be clear. I have more trouble with it on a transactional level because it pushes a certain reading of the documents and relationships based on the predefined categories and ChartEx's idea of how these relate to each other"*.

TASK 3

Task 3 required participants to identify the dates on which “Aalbert” purchased his land. The majority of participants found this task to be “easy” giving it a mean rating of 4.5 (SD 0.85, range 3-5). Similarly, the participants’ confidence in their responses was very high with a mean rating of 4.9 (SD 0.32, range 4-5).

After searching for “Aalbert” once again and opening the documents in which he appears, each of the participants used the “Document Text Highlighting” and “Show Markup” features to highlight only the “Date” entities within the document text. One participant did not complete this task at all due to a computer problem.

Participants found the “Document text highlighting” feature very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 5.0, SD 0.00, range 5). One participant said of the colour-coded markup scheme that *“the colours are good”*. One participant wondered whether the dates in the document text could be cross-referenced with modern dating “*so two different medieval dating methods referring to the same day could be seen as the same date”*.

Participants found the “Show Markup” feature very easy to use (mean rating 5, SD = 0.0, range 5) and very useful (mean rating 5, SD 0.0, range 5). One participant felt that: *“after you select one option to highlight, it should be preserved when you switch documents”*. An other person felt that the large number of “Person” entities that are displayed in a document *“can be cumbersome”* and suggested that they are *“presented in a list or an alternative format that's more organised visually”*.

TASK FOUR

Task 4 required participants to discover another name by which “Aalbert” might be known in the Cluny collection. All participants found this task to be “easy” or “very easy”, giving it a mean rating of 4.5 (SD 0.52, range 4-5). Similarly, the participants’ confidence in their responses was very high with a mean rating of 4.7 (SD 0.47, range 4-5).

After searching for “Aalbert” once again and opening one or more documents in which he appears, almost all of the participants scrutinised the “Possibly Also…” section of the Entities panel to view calculated relationships between entities. One person searched for different variations of the name “Aalbert” (e.g. “Aalberto, “Alber” etc.) and compared document texts manually. One participant did not complete this task at all due to a computer problem.

Participants who used the “Document Results” feature found it very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 4.8, SD 0.45, range 4-5). One participant suggested that the option to sort the results by date would be “great” but acknowledged that it *“might clutter”* the interface. One participant was surprised that the alternative name for “Aalbert” did not appear in the Document Results.

Participants who used the “Entity Results” feature found it very easy to use (mean rating 4.8, SD = 0.67, range 3-5) and very useful (mean rating 4.6, SD 1.33, range 1-5). One participant commented that the entity highlighting within the Entity Extract snippet was *“great”*. Another pointed out that it *“wasn’t immediately clear that I needed to click on the name itself”*.

Participants found the “Document Text” feature very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 4.8, SD 0.67, range 3-5).

Participants found the “Show Markup” feature very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 5, SD 0.0, range 5). One participant commented that the feature was *“very useful when looking for specific information”*.

Participants did not find the “Transactions Visualisation” feature very easy to use (mean rating 3, SD 1.73, range 1-4) and did not find it very useful (mean rating 3.3, SD 2.08, range 1-5). One participant commented: *“It needs getting used to, but can be helpful”*. Another said they were *“aware of it, but not sure how it would have helped”*.

Participants found the “Possibly Also...” feature very easy to use (mean rating 4.4, SD 0.73, range 3-5) and very useful (mean rating 4.6, SD 0.52, range 4-5). Two participants pointed to problems in the presentation of this section, with one not realising that it was an expandable tree diagram that revealed more information. The participant did however acknowledge that *“the arrows are fairly clear markers for additional information”*. Similarly, another participant felt that the tree should be fully expanded by default: *“It might be easier if the name is shown already, without having to click on the document first”*. One participant felt that some of the relationships between entities were missing, commenting: *“Why does Adelbar only appear as a possibility in only one of the Aalbert documents?”* Another participant said it was *“a bit unclear that every person appears multiple time as person”*.

Participants found the “Person Visualisation” easy to use (mean rating 4.3, SD 0.76, range 3-5) but did not find it particularly useful (mean rating 3.7, SD 0.76, range 3-5). One participant commented: *“I don't find this that useful, since the “Possibly Also…” section also has the same information and is easier to understand”.* Conversely, one participant said: *“I ignored the “Same As…” section and went straight to this”*. Another participant said: *“For this exercise, it didn't seem that necessary. I'm not sure when I would need it, but I'm sure if I worked with it more, I would appreciate it”*. Another participant simply asked, *“Why circles?”*

TASK FIVE

Task 5 required participants to identify the name of the wife of a person called “Ugono”, who appears in the Cluny collection. All participants found this task to be “very easy”, giving it a mean rating of 5 (mean rating 5, SD 0.0, range 5). Similarly, the participants’ confidence in their responses was very high with a mean rating of 4.8 (SD 0.41, range 4-5).

Almost all of the participants used the “Entity Extract” feature in the “Entity Results” to complete the task. Three participants opened the relevant documents and used the “Document Text” feature and “Possibly Also…” feature to identify the name. One person, noting variations on the spelling of “Ugono” entered the partial word “Ugon” into the search field. One participant did not complete this task at all due to a computer problem.

Participants found the “Entities to search for” feature very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 5, SD 0.0, range 5). One participant said *“It would be useful if the search could detect different Latin definitions of words, especially occupations”*.

Participants found the “Document Results” feature very easy to use (mean rating 4.8, SD 0.41, range 4-5) and very useful (mean rating 4.8, SD 0.41, range 4-5).

Participants found the “Entity Results” feature very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 5, SD 0.0, range 5). One participant felt the Entity Results were *“very useful in distinguishing similar names when looking for persons, rather than dates or institutions”*.

Those participants who used the "Document Text" feature found it very easy to use (mean rating 5, SD 0.0, range 5) and very useful (mean rating 5, SD 0.0, range 5).

Those participants who used the "Show Markup" feature found it very easy to use (mean rating 5, SD = 0.0, range 5) and very useful (mean rating 5, SD = 0.0, range 5). One participant commented that this feature was *"particularly useful"*.

Only one participant used the "Transactions Visualisations" feature for this task. They felt it was very easy to use (rating 5) and very useful (rating 4).

Only three participants used the "Same As..." and "Possibly Also…" features in the Entities panel. They found it very easy to use (mean rating 5, SD 0.0, range 5) but not very useful (mean rating 3.7, SD 1.53, range 2-5). One participant said the feature was *"not that useful, as many of the name variants did not appear"*. Another participant noted *"If I click on "Lillie" it gives Lillia with low probability but not Lilisa, but if I click on Lilisa, it gives Lillia but not Lillie"*. The participant suggested: *"Maybe there is a way to combine them, so both show up as possibilities?"*

Only one participant used the "Person Visualisation" feature for this task. They felt it was very easy to use (rating = 5) but not very useful (rating = 3).

B) OVERALL MEASURE OF USER EXPERIENCE

The overall measure of user experience comprised 21 items designed to measure 4 different aspects of user experience: *perceived ease of use*, *perceived usefulness*, *disorientation*, and *aesthetic quality*. Perceived ease of use refers to the extent to which an individual believes that using the application will be free of effort. Perceived usefulness refers to the extent to which users perceive that using the application in their job will increase their job performance. Disorientation refers to the feeling experience by users who do not know where they are within the application or how to move to desired locations. Aesthetic quality refers to properties of the application associated with its visual appeal.

PERCEIVED EASE OF USE

Table 1 (below) shows the different items used to measure perceived ease of use. Participants rated each item on a scale from 1 to 7, where 1 = strongly agree and 7 = strongly disagree. The overall measure of perceived ease of use was calculated from a composite score of the 3 items.

The results suggest that participants found the ChartEx Virtual Workbench application relatively easy to use, with the overall mean rating *below* the midpoint of the scale.

Table 1: Descriptive statistics for perceived ease of use

	N	Minimum	Maximum	Mean	Std. Deviation
Learning to use this application was easy	11	1	7	2.7	2.00
Becoming skillful at using this application was easy	11	1	7	3.1	2.26
The application was easy to navigate	11	1	7	2.9	2.07
Perceived ease of use (overall)	**11**	**1**	**7**	**2.9**	**2.09**

PERCEIVED USEFULNESS

Table 2 (belowabove) shows the different items used to measure perceived usefulness. Participants rated each item on a scale from 1 to 7, where 1 = strongly agree and 7 = strongly disagree. The overall measure of perceived usefulness was calculated from a composite score of the 4 items. The results suggest that participants found the ChartEx Virtual Workbench application relatively useful, with the overall mean rating *below* the midpoint of the scale.

Table 2: Descriptive statistics for perceived usefulness

	N	Minimum	Maximum	Mean	Std. Deviation
Using the application would improve my performance in my work	11	1	7	3.1	2.51

Using the application in my work would increase my productivity	11	1	7	3.3	2.24
Using the application would enhance my effectiveness in my work	11	1	7	3.4	2.20
I would find the application useful in my work	11	1	7	3.1	2.39
Perceived usefulness (overall)	**11**	**1**	**7**	**3.2**	**2.27**

DISORIENTATION

Table 3 (below) shows the different items used to measure disorientation. Participants rated each item on a scale from 1 to 7, where 1 = never and 7 = always. The overall measure of disorientation was calculated from a composite score of the 7 items. The results suggest that participants did not find the ChartEx Virtual Workbench application disorientating, with the overall mean rating well *below* the midpoint of the scale.

Table 3: Descriptive statistics for disorientation

	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
I felt lost	11	1	3	1.8	0.60
I felt like I was going around in circles	11	1	4	1.6	1.03
It was difficult to find a page that I had previously viewed	10	1	6	1.5	1.58
Navigating between pages was a problem	11	1	2	1.4	0.50

I didn't know how to get to my desired location	9	1	3	1.9	0.60
I felt disorientated	9	1	3	1.7	0.71
After browsing for a while I had no idea where to go next	9	1	2	1.3	0.50
Disorientation (overall)	**11**	**1**	**3.1**	**1.6**	**0.61**

AESTHETIC QUALITY

Table 4 (below) shows the different items used to measure aesthetic quality. Participants rated each item on a scale from 1 to 7, where 1 = the negative extreme of the item and 7 = the positive extreme of the item. The overall measure of aesthetic quality was calculated from a composite score of the 7 items. The results suggest that participants found the aesthetic quality of the ChartEx Virtual Workbench application to be positive, with the overall mean rating *above* the midpoint of the scale.

Table 4: Descriptive statistics for aesthetic quality

	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
I judge the application to be very complex ~ very simple	9	2	6	4.1	1.27
I judge the application to be very illegible ~ very legible	9	5	7	5.9	0.60

I judge the application to be very disordered ~ very ordered	9	5	7	6.2	0.67
I judge the application to be very ugly ~ very beautiful	9	4	7	5.1	1.05
I judge the application to be very meaningless ~ very meaningful	9	6	7	6.3	0.50
I judge the application to be very incomprehensible ~ very comprehensible	9	5	7	6.0	0.71
I judge the application to be very bad ~ very good	9	5	7	6.6	0.73
Aesthetic quality (overall)	**9**	**4.9**	**6.6**	**5.7**	**0.50**

6. DISCUSSION

This evaluation has demonstrated through both the ratings and comments from participants that their perceptions of the ChartEx Virtual Workbench are extremely positive. Participants were able to use the application to easily complete each of the tasks. The application also gave them confidence that the answers they found were correct.

The participants varied in their approach to most of the tasks, demonstrating the versatility of the ChartEx Virtual Workbench in allowing different "routes" through the data. With the exception of the two visualisations used in the application, participants found all of the features both very useful and very easy to use.

The *"Collections to search"* and *"Entities to search for"* features were used in the majority of tasks, to narrow the search space. Whilst participants generally found these both useful and easy to use, they identified a number of things that would add further clarity. These included: providing explanatory text about each the collections and the entity types, distinguishing between (or merging) similar entity types, and explaining the purpose of the colour coding.

The *"Document Results"* and *"Entity Results"* features were also used in the majority of the tasks, to display the search results. Again, participants found these both useful and easy to use. Participants found the search results particularly useful for displaying variations in the spelling of entities but criticised the inability to order the results by date.

The *"Document Extract"* and *"Entity Extract"* features proved particularly popular with participants, and were used in the majority of tasks. Whilst it was intended that the extracts would provide only a brief preview of relevant documents, many participants were able to complete the tasks using these features alone. Indeed, the length of the extracts was mentioned by a number of participants. Some felt the extracts were too short, others felt they were too long, and some felt the length of the extract should be adjustable. Also, some participants felt the difference between the *"Document Extract"* and *"Entity Extract"* features was not clear enough.

The *"Show markup"* and *"Document text highlighting"* were used in conjunction across many of the tasks, to toggle and display the entities marked up in the documents. Again, participants found these features both useful and easy to use. Whilst the colour scheme was praised by some participants, others found the amount of coloured highlighting on some documents to be overwhelming and would prefer it to be disabled by default. Others wanted selected entity types to be preserved across documents rather than having to select/deselect them for each new document opened.

The *"Same As…"* and *"Possibly Also…"* features in the Entities panel were used in some of the tasks, to display both confirmed and calculated relationships between entities. Participants found these both useful and easy to use but criticised the expandable tree diagram used to display the information. A number of participants were unaware that each branch of the tree diagram could be expanded, which resulted in them failing to find the information. Other participants identified gaps in the data, which resulted in some relationships only partially being displayed.

The only features that were not rated positively by participants were the *"Transactions Visualisation"* and *"Person Visualisation"*. Participants did not find them very easy to use and did not find them very useful. This was largely due to the design of the visualisations, which used concentric circles to represent connections between entities. Participants felt the purpose

of the circles was unclear and struggled to relate them to the data. Some participants felt the limited amount of space available to display the visualisations made them difficult to use, and others felt the visualisations duplicated information available elsewhere in the application. One participant felt that the visualisations pushed a particular reading of the documents and the relationships within them.

The overall measure of user experience indicated that the participants found the ChartEx Virtual Workbench application relatively easy to use and relatively useful. They did not find the application disorientating and found the aesthetic quality of the ChartEx Virtual Workbench application to be positive.

This evaluation has demonstrated that the ChartEx Virtual Workbench is useful, easy to use and supports the exploration of medieval charters. The participants identified a number of improvements that could be made to the application in future but overall, both historians and other user groups received it positively.

7. REFERENCES

Beyer, H. and Holtzblatt, K. (1997). *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann Publishers Inc., San Francisco.

E. CONCLUSION

1. ChartEx did develop and deploy a system that combines NLP and DM to extract data useful to researchers regarding locations and related actors, events and dates from digital charters.
2. The evaluation of ChartEx also suggests that the virtual workbench based on novel instrumental interaction techniques does produce more useful knowledge from charters than a human working alone or an automated NLP/DM system working alone .
3. The ChartEx system can produce efficient and accurate knowledge from charter documents by processing the information in English summaries of the charters. The

work on Latin charters is not completed. The results of the Data Mining workpackage have so far been restricted to manually annotated samples.

4. The historians within ChartEx did develop an ontology for the investigation of both Latin charters of UK provenance and Latin charters from France.

F. DISCUSS HOW YOUR PROJECT PROGRESSED OVER TIME, AND HOW YOU MANAGED IT

In the initial phase of the project we developed a project plan:

[INSERT Project Plan diagram – submitted to JISC - here]

After a launch meeting at TNA in London in January 2012 the first five months of the project focussed on developing an ontology to support the manual markup of training sets of charters from the selected collections. Thereafter the idea of a progressive project plan in which the different components built progressively on each other was replaced by a system in which each of the three core components (NLP, DM and the VWB) developed in parallel using the manually annotated datasets as exemplary data. In addition partner at University of Washington experimented with developing a processing solution using LOD (see appendix three).

Collaboration was managed through a series of meetings (both face-to-face and skype) and by the use of an online management system for project records.

G. EVIDENCE HOW YOUR PROJECT HAS IMPROVED THE RESEARCH ENVIRONMENT;

Evidence of success in improving the research environment is provided by funding for further projects:

1) AHRC Big Data (Arno Knobbe and Roger Evans, with TNA)....
2) DID3 Helen Petrie and Chris Power.....

In addition the Borthwick Institute for Archives and the Department of History at the University of York are investing in some further work in order to evaluate the results of ChartEx, among other projects, towards developing next-generation archival training.

H. DOCUMENT MEETINGS AND IMPORTANT MILESTONES;

Meetings were documented on the project online management site (Basecamp) and conference presentations are published at www.chartex.org. Milestones are indicated in the research reports above.

I. DESCRIBE LESSONS LEARNED (BOTH POSITIVE AND NEGATIVE);

Positive advantages all round in approaching the problem from a new perspective (beyond established digital humanities approaches and systems). The advantage of tackling innovative solutions.

Importance of goodwill, and good communication skills, combined with flexible advanced scheduling to sustain international collaboration. Virtual meetings need to be backed up through face to face meetings.

Need for independent arbitration in technical disputes. Need for clarity about the degree of integration expected from Digging into Data Projects which lack the overall project structure and hierarchies of (for example) EU projects. Different national funders had very different reporting expectations.

J. DOCUMENT ANY SOFTWARE (INCLUDING IP ARRANGEMENTS), ALGORITHMS, OR TECHNIQUES THAT YOU DEVELOPED AND HOW THESE MIGHT BE SUSTAINED OVER TIME;

See above under 'Results of Research' and below in Appendices 1, 2 and 3.

IV. APPENDIX ONE CHARTEX INITIAL MARKUP SCHEMA: GUIDELINES

A. 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 should take a back seat to the meaning of the text itself, even if that allows for greater ambiguity. This will be reinforced below; see in particular the **What to Mark Up?** section.

B. INVERSE RELATIONSHIPS IN MARKUP

Development of the markup schema has led us to the conclusion that "is" constructions are more instinctive than "has" constructions, and this is reflected in the markup. As an example, there was a consensus that:

- Josce is father of Thomas

Is preferable to:

- Thomas has father Josce

In addition, it is only necessary to mark up relationships between entities in one direction – the computer can extrapolate the other half of the equation. So if you mark up "Josce is father of Thomas" there is no need to also mark up "Thomas is son of Josce". It doesn't matter which you choose to specify, although for some relationships only one of the paired relationships is available in the mark up tool, so the decision will be made for you.

C. ENTITIES: ACTORS, SITES, EVENTS, ATTRIBUTES

A more detailed description of the various entities that we have selected to be marked up follows. They are presented in the order in which we suggest they be dealt with (i.e. do **Document** and **Apparatus** first, then actors, then move on to **Sites** and so on), although there is technically no right or wrong sequence in which to approach them.

D. DOCUMENT

The **Document** is of vital importance, and may vary widely. It is simply a unique number assigned to the document, which allows us to place it in context and retrieve such information as provenance and publication data. It may range from an archive number (as with the Borthwick material) to a number assigned by a previous database (as with the DEEDS material).

E. APPARATUS

The **Apparatus** entity is used solely to mark up the work of an editor. This can include notations as to endorsements, for example, or may encompass truncations of portions of the original text. For example:

John grants his tenement (location given) in Bromley to Colchester Abbey.

Here (location given) should be marked up as **Apparatus**, as it indicates that an editor has excised a portion of the original document.

Dates assigned by an editor should be marked up using **Apparatus**, NOT using the **Date** entity below.

F. ACTORS – PERSONS, INSTITUTIONS, AND ACTORS

Actors are people (named or not), ecclesiastical or secular Institutions, or offices. Examples, respectively, might be: Robert fitzHugh, Glastonbury Abbey, and Dean of York. All such actors in a document should be marked up, even if their exact function is unclear or unknown (how to do this is addressed below). The distinction between unnamed people and offices is also outlined below.

Persons

A **Person** is, essentially, an individual actor. In typical legal document, a **Person** may be acting on their own behalf, or in consort with an **Institution** with which they are affiliated. To distinguish a person acting in consort with an Institution from a specific institutional office, the presence of a personal name is key. For example:

Abbot John and the convent at Colchester grant land to a priory.

Here Abbot John, because he is named, is a **Person**. See also **Institutions**, below. He is also, obviously, a member of the Institution 'the convent at Colchester' – this relationship is not overlooked, and is addressed in the linking of entities which define their **Roles** (see below).

Note that his title, "Abbot" should be considered part of his name. It will also be marked up as his **Occupation** (see below). This is true of all titles appearing before individual names, though generic ones such as Brother will not be considered **Occupations**.

On occasion we may see individual actors who are not named and who do not fit into the category of **Institutions** below. Such instances would typically be subsequent mentions of a previously named **Person** or abbreviated mentions of a **Person**. For example:

The said R grants in exchange a certain field called Topfield.

Here we know R's full name from an earlier point in the document, but he would also be identifiable as a Person even if we were only ever given an initial. Another example:

Robert fitzHugh gives to his eldest son a certain toft.

Here Robert's "eldest son" is also an actor and a **Person**, and will be marked up as such, although we do not know his name.

On Surnames:

This is a question that has vexed us considerably: when do we call something a surname and when is it simply a description of an individual? As the answer to this question varies depending on time and Place, we have decided upon the following as a possibly imperfect but working approach.

-A surname is part of the entity **Person** and a first name and surname should be marked up as one.

-Allow the text to determine what is and is not a surname, *even if this results in inconsistencies within a document or between multiple documents.* For example, "Thomas of York, smith" may well be the same Person as "Thomas Smith of York". It is up to the computer to suggest links between two such **Persons** if they appear in separate documents; within a single document they may be linked using the role "is same as" (see the **Roles List** below).

-There are a number of possible surname constructions: patronyms (e.g. fitz Hugh); toponyms (e.g. of York); and occupational names (Smith). In each case, we follow the documents' lead, as above. The portion following immediately after the first name is considered the surname and together they identify the **Person.**

-We distinguish between John Smith and John the Smith in English. Names using Anglo-Norman "le" are surnames. Similarly, distinguish between Robert fitz John (surname) and Robert son of John in edited summaries. In Latin documents we follow our surname principle rigorously, meaning that Adam filius Willelmi is all one name.

-There should be NO embedded markup concerning the surname of a **Person** (e.g. for "Thomas Smith", "Smith" should NOT also be marked up as an **Occupation;** for "Thomas fitz Hugh", "Hugh" should NOT also be marked up as a separate **Person**).

-Some examples (the underlined portion is the **Person)**:

Thomas Smith of York, son of Hugh

Thomas of York, smith, son of Hugh

Thomas fitz Hugh, of York, smith

The additional descriptors of the **Person** following the surname should be treated separately as **Places, Occupations,** and other **Persons.** Their relationship to the **Person** will be reflected in the connections made through **Roles.**

-In edited English translations, we follow the editor's lead. Where the editor has distinguished between *Thomas le Spicer* (a **Person**) and Thomas the spicer (a **Person** + **Occupation**), we do as well. In Latin documents we follow our above guidelines more doggedly.

Institutions

An **Institution** is, in contrast to a **Person**, a corporate actor. In most cases, an Institution is easily identifiable according to modern concepts of what constitutes a secular or ecclesiastical **Institution** (e.g. churches, abbeys, deaneries, etc.). Because land could pertain to whomever happened to be holding a certain position within an **Institution**, specific offices associated with **Institutions**, if they are not attached to a given personal name, are also considered **Institutions** for our purposes. For example:

The Prior of Taunton is to receive rent from certain properties nearby.

Here the prior is an **Institution**, because the income pertains to the office of prior, not to a specific man who holds it. The income will pass to the next prior, and will not attach to any particular individual after they have left the office.

Actors

An **Actor** is an entity used only in the rare circumstance that it is unknown whether a participant in a **Transaction** is (or will be) a **Person** or an **Institution**. For example:

Bishop Richard grants the income from certain land to whomever shall hold the mill at North Curry.

Here "whomever" is an **Actor**, as the mill could be held by either a **Person** or an **Institution**. Note however that such cases will be quite atypical.

G. LOCATIONS – SITES AND PLACES (AND PARCELS)

Locations are the basic unit with which we are concerned. The purpose of this markup is to highlight and extract information connected with Sites by making visible the web of relationships between them and the surrounding world. Thus identification of these geographical units is crucial. Any given document may contain multiple locations, or only a single one.

The larger category of locations is subdivided in two: **Sites** and **Places**. **Parcels** are an associated concept.

Sites

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 spot; it may be noted by any number of terms, from 'toft' to 'field' to 'ditch' to simply 'land'. It may be the main piece of property discussed by a **Transaction** (see **Parcel**, below), or it may be a smaller Site specified as within that main Site. It may function as a boundary marker delimiting length or width, or it may be mentioned as a directional marker, demarcating (for example) the northernmost limit of another Site. Any

geographical spot mentioned in a document may safely be marked up as a Site once it has been determined that it is not a **Place** (see below).

Multiple **Sites** such as streets, fields, parks, cemeteries, schools, private properties can together make up **Places** (see below). **Sites** are usually clearly bounded with visible or marked boundaries (even if these are disputed or unknown). They may be in common ownership (e.g. streets) or in private ownership (houseplots).

Some **Sites** may have proper names (as is more typical of **Places**) – streets, fields, parks, inns, religious buildings typically have proper names, and so do some houses. Technically these are 'urbonyms'. Scholars of toponyms (see **Places** below) often collect and study urbonyms as well (this is true of the English Place Name Survey now being digitised by DEEP). This blurs the distinction being made here between **Place** and **Site**, but for the purposes of ChartEx we will maintain the distinction.

SiteRef

SiteRef is an entity designed for use when a **Site** is mentioned only as a pronoun. For example:

John gives Richard the tenement between the land of William and that of Adam.

Here "that" is the only word that can represent Adam's land, and it must be marked up. If Adam's land is mentioned as a noun elsewhere in the document (and it is evident from the text of the document itself that "that" is unambiguously the same location),"that" does NOT need to be marked up as the relationships can be attached to the noun. SiteRef, like **PlaceRef** below, should only be used for pronouns that do not have antecedents. In this sense it serves a similar function to **Actor**, and will probably be quite rare.

Places

Places are distinct from **Sites** in that they are larger, geographically delineated, named conglomerates of many **Sites**. A typical **Place** might be a county, hundred, diocese, city, town, parish etc., but in smaller rural communities, vills and manors are **Places**.

Place-names are usually referred to as toponyms (see the discussion on urbonyms under **Sites,** above).

They are not always clearly bounded (e.g. Exmoor). When they are bounded by administrative boundaries, these may not be visible and/or may not be consistent with each other (ecclesiastical boundaries may not coincide with secular administrative boundaries). This inconsistency contributes to uncertainty about the extent of a **Place**. Where is York – is it the

walled city, the historic city, the contemporary unitary authority? Different people will answer this question differently.

Places can be nested. Kew is a **Place** but it is also part of London, which is a larger **Place**. The parish of St Michael is a **Place** within the City of York (also a **Place**). **Places** contain many **Sites** (even many thousands of **Sites**).

PlaceRef

PlaceRef is an entity designed for use when a **Place** is mentioned only as a pronoun. For example (rather laboured, but you get the idea):

John gives Richard the wood between the town of Barby and that to the east.

Here "that" is the only word that can represent the town east of Barby, and it must be marked up. If the town is mentioned by name elsewhere in the document (and it is evident from the text of the document itself that "that" is unambiguously the same location), "that" does NOT need to be marked up as the relationships can be attached to the noun. **PlaceRef**, like **SiteRef** above, should only be used for pronouns that do not have antecedents. In this sense it serves a similar function to **Actor**, and will probably be quite rare.

****Parcels**

Parcels are NOT marked up as entities like **Sites** and **Places** are. Rather, the concept of parcels comes into play when we are defining **Site-Transaction Roles** (see below). However, it is discussed here as it is perhaps easiest to conceptualize in the context of **Sites** and **Places**.

A parcel is the main piece of property discussed by a **Transaction**. It may or may not be what is conveyed by the **Transaction** itself; some **Transactions** transfer a **Site**, while others transfer rights or income associated with a **Site**. This distinction as to what specifically is being conveyed is not relevant to our markup at this stage, but we are interested in picking out the main Site associated with each **Transaction**.

As suggested by the above, a parcel is typically a **Site**, although in certain rare instances it can be a **Place**. A parcel may contain smaller **Sites** and may be described using additional **Sites**. A document may contain more than one parcel, especially if it contains multiple **Transactions**.

For example:

Bishop John grants income from land in the hundred of Wells to Glastonbury Abbey, namely from the woods called the Grava bounded to the north by a stream and to the south by Robert son of Adam's farm, and from….

Here ‘land’ is a **Site**, and its role is that of a parcel in the **Transaction**. The ‘hundred of Wells’ is a **Place**. The ‘woods called the Grava’ is a **Site**, as is the ‘stream’ and the ‘farm’.

Sometimes a **Transaction** may refer to multiple **Parcels.** It can be tricky to know how to group **Sites** and **Places** in relation to their role as **Parcels.** For example:

Bishop John grants income from a house with garden and from two acres of meadow…

Here ‘house’, ‘garden’ and ‘two acres of meadow’ are all sites. ‘Garden’ would best be considered as *part of* ‘house’ (see the roles list below on *is part of* relationships), while both the house and the two acres of meadow should be **Parcels** in the transaction ‘grants’. Ultimately, the decision as to what should be a separate **Parcel** vs what should be *part of* the **Parcel** rests with the person doing the mark up. Note that fewer parcels can help minimise clutter in a marked-up document.

H. EVENTS – TRANSACTIONS, DATES, AND EVENTS

Events are our temporal clues for the mapping of Sites over time. They can include **Transactions**, any **Dates** associated with the document, and specific, datable **Events** such as fires or episcopal or monarchical reigns.

Transactions

Transactions are the acts that are recorded in the documents, and they may take many forms: grants, exchanges, quitclaims, confirmations, etc. We consider them to be Events as they are in fact legal actions at a particular moment in time. Our main concern with these is not to focus on the legal function of the document or the exact nature of the conveyance, but to keep our attention on how **Transactions** relate to **Sites**. A **Transaction** is identified as the main verb that describes the act in question, where the act in question in understood to involve **Sites** or **Places** in some way. There may be multiple **Transactions** in a single document. For example:

Abbot Richard quitclaims rent pertaining to certain fields to Simon Montacute; for this quitclaim Simon grants to the Abbey the income of the mill at Duck’s Corner.

Here there are two **Transactions**: the quitclaim and the grant, each with a different **Site** functioning as a parcel (certain fields and the mill, respectively).

Note that only **Transactions** explicitly involving Sites or Places should be marked up. You may encounter things like warranty clauses, or details of the terms of payment of rent. If there is no mention of a geographic location in the clause, it is not useful for our project. Furthermore, if there is no *new* geographical information in a clause that mentions locations already discussed (such as in a clause detailing terms of payment of rent, for example), it should not be marked up. Similarly, any clauses involving hypothetical circumstances (for example, a restitution or

distraint clause) should not be marked up. Thus there may be any number of clauses that are vital to the legal shape of the document, which we will be omitting from our markup.

Agreements, exchanges, notifications, and final cocords are types of documents that often include multiple **Transactions** of equal importance. At the moment our approach is to focus on the **Transactions** which comprise the exchange, agreement, etc. Thus, given:

It was agreed between John fitzWilliam and Bishop Richard that John grants Bishop Richard land in Smithfield and Bishop Richard quitclaims to John his rights and claim in two tenements.

> -We do not mark *agreed* as a **Transaction**. The same applies to other verbal phrases opening similar types of documents.
>
> -Both *grants* and *quitclaims* are **Transactions**, as they both involve **Locations.**
>
> -The **Document ID** refers to both **Transactions**. There is no need to link them using "same as" or "not same as". If both are of equal weight and both involve **Locations**, any witnesses should pertain to both. If one is obviously of less weight or does not refer to **Locations**, the witnesses can pertain only to the main **Transaction.**

Dates

Dates may appear internally, as dating clauses, or externally, as assigned by editors or archivists. If they are assigned, they should be marked up using **Apparatus** (see above). They may be specific to the day, or may loosely cover a period of many years. In the case of internal dating clauses, a wide range of dating styles may apply. Documents will likely include everything from the Julian calendar to regnal dates of both ecclesiastical and secular officials to saints' days and papal indictions. All are acceptable and should be included.

Events

Events are occurrences that are specific and notable enough that they can help us Place changes pertaining to Sites in some kind of temporal order. A large fire in a city, for example, may be mentioned in a document as a point of reference. Other Events might include details pertaining to certain individuals' tenure in office, such as a royal visit or an episcopal inspection. For example:

Since the Monastery of Gisburn was recently burned along with all its contents, we are moved to grant five acres of land towards the rebuilding efforts of the monks.

Here 'the Monastery of Gisburn was recently burned' is an **Event**, which precedes the **Transaction** (in this case, a grant).

I. ATTRIBUTES – OCCUPATIONS

Unlike other entities, the purpose of attributes is really to clarify specific information about a single entity by giving it a tagged piece of text to link to in a set, immutable way. This is in contrast to the other entities, which link to one another in a variety of relationship-defining ways. That said, this does not fundamentally change the process by which markup of attributes occurs.

Occupations

Occupations relate solely to **Persons**, and may include a large number of 'jobs', such as ecclesiastical positions (abbot, dean, clerk), governmental positions (justiciar, sheriff, bailiff), and trades (smith, wheelwright, blacksmith). Occupational information may elide with surnames, especially in earlier periods, or with titles related to Institutional positions. See the section on **Surnames** above. For example:

John Faber receives land from Bishop Simon.

Here 'John Faber' is a **Person**, but 'Faber' is his surname and should not be marked up as an **Occupation**. 'Bishop Simon' is also a **Person**, and 'Bishop' is also his **Occupation.** Here there is some overlap of the markup.

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.

When faced with *Simon precentor of York* our approach is as follows:

Simon (person) *precentor of York* (occupation)

Unless an institution like a church or abbey is specifically mentioned, we group the individual position together with the of place/site as the occupation. So, as a counterexample, if we had *Simon precentor of the church of York*, we would mark it up as:

Simon (person) *precentor* (occupation) *York* (institution), with Simon an officeholder in York

J. WHAT TO MARK UP?

The following is a quick run-through of exactly which parts of the document are to be tagged for markup when considering each entity as discussed above. It is important that tagging be applied consistently across the project. Note that the software we are using allows markup to overlap.

- Note: A basic principle guiding our application of mark up is to always aim for the smallest unit of text that will do the job. Remember that we want to draw out the relationships between actors, locations, and Events, so there is no appreciable advantage to marking up "the croft of Adam fitz John in Bell's lane" as one **Site**. Rather, mark up "croft" as a **Site**, "Adam fitz John" as a **Person**, "John" as a **Person** (for more on how to identify a surname and how to mark up "fitz" surnames, see the discussion **On Surnames** under **Persons** in the section on **Entities** above), and "Bell's lane" as a **Site**. The **Roles List**, below, clarifies how to connect each of these entities to one another through relationships that define their role in the document.

1. DOCUMENT

Mark up:

- Some sort of identification number or archival reference will introduce each document in the ChartEx repository. This is the **Document**. It should appear at the top of the document.

Don't mark up:

- Additional information about provenance or editions; this should come under **Apparatus**

2. APPARATUS

Mark up:

- Any emendations, truncations, annotations, or additions made by an editor. This may include assigned dates or date ranges. This may include vocabulary from the original Latin text if you are working on an English translation. If it appears in brackets, it should be seen as an editorial interjection and thus it is **Apparatus.**
- A basic principle to guide you: Text should only be marked up as **Apparatus** if it is interpolated material – i.e. if it were extracted the contents of the document would not be appreciably altered. This means that sections marked up as **Apparatus** should not contain any embedded markup of other entities.
- Mark up **Apparatus** as a single entity as much as possible. There is nothing to be gained from creating separate segments, and this helps reduce visual clutter in the mark up tool.
- **Apparatus** should not contain any other markup embedded within it. However, other entities can contain embedded **Apparatus** markup. For example, *William [de Aurelianis]* – the whole can be the **Person**, and *[de Aurelianis]* selected within that as **Apparatus**.

Don't mark up:

- Some of the documents we are dealing with are, in a sense, composed entirely of **Apparatus**, as they are English summaries of original Latin documents. For our

purposes, treat the main body of the document as a base text and only treat very obvious editorial interventions as **Apparatus**.

3. ACTORS – PERSONS, INSTITUTIONS, AND ACTORS

Persons

Mark up:

- Personal names and surnames. For more on how to identify a surname and how to mark up patronyms, toponyms, and surnames derived from occupations, see the discussion **On Surnames** under **Persons** in the section on **Entities** above.
- Very generic titles like Master, Brother, Lord, etc. should be marked up as part of the **Person**, not as an **Occupation**.
- Titles indicating that the individual holds a particular office, like Prior or Sheriff, should be marked up both as part of the **Person** and as an **Occupation** (see below), *if* they appear together with the individual's name (e.g. "Prior John" is a **Person**; "Prior" is also the **Occupation.** But for "John, prior of Waltham Abbey", "John" alone is a **Person**).
- Subsequent mentions of a **Person** using their name or first initial (i.e. as a proper noun), even in clauses and **Transactions** that are not themselves being marked up. Similarly, first mentions of **Persons** in insignificant clauses should be marked up and linked whenever possible.
- Individuals appearing in the witness list, who will be linked to the **Transaction** even though they may have little or nothing to do with the **Sites** central to it. This is because witness lists can provide valuable context for dating such documents and/or placing them in a broader community of those involved with land **Transactions**.

Don't mark up:

- Relative clauses providing more information about an individual. In *John, son of the miller Adam*, for example, John and Adam should be separate **Persons**.
- Personal pronouns referring to already mentioned **Persons** (i.e. with an antecedent) can safely be omitted. See also **Actors**, below.
- Justices (itinerant or otherwise) appearing in royal courts have little bearing on the relationships between **Locations** and **Actors** we are interested in; therefore we **do not mark them up**. This is a special case, and one which you will typically find in final concords and similar documents.

Institutions

Mark up:

- Corporate actors, which may be obvious, like the "Church of St John", or less obvious, like the "Prior of Taunton."
- A rule of thumb: if no personal name is provided along with the title of the office, it is most likely that you are looking at an **Institution**.

- Where an office of a particular institution is mentioned, the whole thing is a single **Institution**. For example, the "dean of the church of York" is just one **Institution,** not two linked together.
- Subsequent mentions of an **Institution** using its name (i.e. as a proper noun), even in clauses and **Transactions** that are not themselves being marked up. Similarly, first mentions of named **Institutions** in insignificant clauses should be marked up and linked whenever possible.

Don't mark up:

- In the case of "John, Prior of Taunton", the actor here is John (a **Person**), whose **Occupation** is Prior of Taunton.

Actors

Mark up:

- ONLY cases where the actor could be either a **Person** or an **Institution**.

Don't mark up:

- Pronouns with antecedents. These can be omitted from the markup –use the antecedent to construct the relationship reflected in the use of the pronoun. For example:

John grants land to Simon. He also grants more land to William.

Here "He" does not need to be marked up. "John" can be used to build links to both Simon and William as a grantor.

4. LOCATIONS – SITES AND PLACES (AND PARCELS)

Sites[/SiteRef]

Mark up:

- Any noun referring to a location as defined in **Sites** above. For Sites that are referred to only by a pronoun without an antecedent (or by numbers serving a similar function), use **SiteRef.**
- Again, keep the markup to the smallest unit of text possible (see the opening comments to this section). However, note that some **Sites** have names that could concievably be split into two (e.g. Snaith ings), but can also be considered a single site. Try to distinguish, as much as possible, between Snaith Marsh (a single site with a proper name) and "a marsh in Snaith" (a site in a place).
- However, we want to include quantities, so "two houses" is a **Site**, as is "three acres of meadow".
- Subsequent mentions of a **Site** using its name (i.e. as a proper noun), even in clauses and transactions that are not themselves being marked up. Similarly, first mentions of named **Sites** in insignificant clauses should be marked up and linked whenever possible.

- On the particularly tricky case of roads: the typical example sees a road as a boundary to a location, with the road itself further specified as leading to x or leading from x to y. We have decided to connect x and y to the road using *is boundary to.*

Don't mark up:

- Be careful not to confuse **Institutions** as actors with **Sites**: if an Institution is playing an active role in the document, it should not be marked up as a **Site** (barring very rare instances where an **Institution** may act in a document as both actor and **Site**)
- Long formulaic phrases that do not reflect any actual **Sites** but are rather present as a form of legalese.

Places

Mark up:

- Any noun referring to a location as defined in **Places** above. For **Places** that are referred to only by a pronoun without an antecedent, use **PlaceRef**.
- Subsequent mentions of a **Place** using its name (i.e. as a proper noun), even in clauses and transactions that are not themselves being marked up. Similarly, first mentions of named **Places** in insignificant clauses should be marked up and linked whenever possible.
-

Parcels

Don't mark up:

- Parcels! This is a concept that comes into play when linking entities to define **Roles** (see below).

5. EVENTS – TRANSACTIONS, DATES, AND EVENTS

Transactions

Mark up:

- The words that enact the conveyance – that is, the verb(s). Remember that we are not marking up diplomatic formulae, so you do not need to seek out a dispositive clause, though this is likely where you will find the verb(s) you want. For example, "grants and concedes" and "quitclaims" are perfectly sufficient.
- Keeping in mind that there are many clauses that we are not interested in here (see **Transactions** above), also remember that a single document may have more than one **Transaction**. For more on how to handle these, see **Transactions** above.

Don't mark up:

- Parts of the document not involving locations or not adding to our information concerning already mentioned locations (e.g. payment clauses, warranty clauses, etc.); parts of the document concerning hypothetical future situations (e.g. distraint clauses, restitution clauses, etc.). We call these "insignificant clauses".
- Long formulaic phrases typical of diplomatic clausulae.

Dates

Mark up:

- **Dates** given in any form, including typical medieval dating practices like regnal years, indictions, feast days, and the Roman calendar. A **Date** may combine several of these into one phrase; that's fine – mark it up as a single entity.

Don't mark up:

- Formulaic phrases commonly appearing along with dating clauses such as "data/facta apud" or "dated at"

Events

Mark up:

- Major occurrences appearing in the document that might contribute to its dating. These might include the death or visitation of a notable figure, a natural or man-made disaster, or a reference to a significant political Event (e.g. the return of King Richard from Crusade or a royal wedding). These will be fairly rare.
- Again, keep the markup to the smallest unit of text possible (see the opening comments to this section).

Don't mark up:

- Any mention of a deceased **Person** (e.g. "quondam Simon") as an event. If the death of a public figure is mentioned as an **Event** in the document, it may qualify. A dead **Person** does not.

6. ATTRIBUTES

Occupations

Mark up:

- Individual words or phrases specifying the **Occupation** of a **Person**, including titles specifying a specific Institutional position such as Prior, Abbot, Sheriff, etc. Such titles will also be marked up as part of the **Person** *if* they appear together with the individual's name in the document text.
- When an **Occupation** takes a form along the lines of "Dean of York" or "archdeacon of Estriding", mark up the whole phrase as the **Occupation.** When a particular **Institution**

is specified, as in "precentor of the church of York", "precentor" should be the **Occupation,** and **"**church of York" an **Institution** to which the **Person** will be linked using "is officeholder in".

Don't mark up:

- Generic titles such as Brother, Master, Lord, etc. These are marked up as part of the **Person** only.
- Similarly, certain broad descriptors such as "citizen" should not be marked up at all.

K. ROLES

"Roles" describe the function(s) of the entities **Actors**, **Sites**, **Events**, and **Attributes**, and thus the connections between them. It is through the roles we attach to each entity that the network of relations in the charter is made apparent. The following is a list of the roles that may be used to connect entities as described above. Remember that we are aiming for uni-directional markup, so once a relationship between two entities has been created, it is not necessary to mirror the relationship in the reverse direction.

The basics:

- If you want to change your mark up of an entity after linking it to another, you have to un-link it first. For this reason, it is worthwhile to mark up all the entities in a document before you begin to deal with their roles in relation to one another.
- Each entity should be connected by at least one role to another (the "Arno" principle)
- A good place to begin is to connect the **Document** to the **Date,** and the **Document** to the **Transaction(s).** This ensures that other relationships (of **Actors**, **Locations**, **Attributes,** and **Events**) can be traced back through the **Transaction** to the **Document** and **Date** and prevents us from having "free-floating" data.
- You don't need to mark up redundancies. For example, if John "is grantor in" a transaction, you do not also need to specify that he "is grantor of" the parcel.
- When marking up relationships between entities, try as much as possible to reflect the language used by the document. For example, given *Johanna wife of Simon*, use the role *is wife of* rather than *is husband of* to connect the two.

Note: In the list below the entities of **Site** and **SiteRef** (and **Place** and **PlaceRef**) have been grouped together. Any role connecting a **Site** can also theoretically be used to connect a **SiteRef**, so the roles list is the same for both. For details on when, why, and how to use **SiteRef** or **PlaceRef**, see the relevant sections above.

A) UNIVERSAL ROLES:

- Entity is same as Entity
- Entity is not same as Entity

Clarifications:

- "is same as" is a symmetric transitive role, which is to say that if A=B=C=D, then A=D. The annotation tool will automatically create the shortest pathway between entities linked in this manner to minimize visual clutter
- We always use *is same as* to connect **Persons**, **Institutions**, and **Sites** to previous appearances of themselves, though without marking them up solely for the purpose of linking them using *is same as* (i.e., we don't need to seek out every use of the word "church" throughout the entire document just to clarify this, especially when it's in a clause that is of little or no interest to us). We only use *is same as* to connect **Places** when there is possible confusion due to orthographic variation.
- However, if you have marked up an **Actor** or **Location** in an insignificant clause (see **Transactions**, above), you may need to connect it using *is same as* so as not to be in violation of the Arno principle, whereby no entity should be completely disconnected.
- is not same as – for use in disambiguating instances when two actors or locations with the same name appear in a document (e.g. John grants land to John). Note, however, that it is not necessary to use *is not same as* very much. If the document contains *terra Reynaldi* and *terra in Snaithing* it is not necessary to disambiguate the two instances of *terra*, as they are already disambiguated through their connections to people or other land.

B) PERSON-ATTRIBUTE ROLES:

- Person occupation is Occupation

C) PERSON-INSTITUTION ROLES:

- Person is officeholder in Institution
- Person is previous officeholder in Institution

Clarifications:

- An example: "John, Prior of Waltham Abbey" - Here John is a **Person,** Prior is his **Occupation**, and Waltham Abbey is an **Institution** in which John is an officeholder.

D) PERSON-PERSON ROLES:

- Person is father of Person
- Person is mother of Person
- Person is parent of Person
- Person is grandfather of Person
- Person is grandmother of Person
- Person is grandparent of Person
- Person is son of Person
- Person is daughter of Person
- Person is a child of Person
- Person is grandson of Person
- Person is granddaughter of Person
- Person is grandchild of Person

- Person is husband of Person
- Person is wife of Person
- Person is spouse of Person
- Person is familial relation to Person
- Person is neighbour of Person
- Person is other relation to Person
- Person is sibling of Person
- Person is brother of Person
- Person is sister of Person
- Person is widow of Person
- Person is widower of Person

Clarifications:

- is familial relation to – intended to cover all other blood relatives (e.g. uncle, aunt, cousin)
- is other relation to – intended to cover all other non-family ties to people (e.g. servant, gardener, attorney)
- where possible, the gendered version of a role is preferable and sufficient (e.g. "is brother" is preferred to "is sibling", and you do not need to mark up both)

E) PERSON-SITE[/SITEREF] ROLES:

- Person is a landlord of Site[/SiteRef]
- Person is a tenant of Site[/SiteRef]
- Person is a previous landlord of Site[/SiteRef]
- Person is a previous tenant of Site[/SiteRef]
- Person is a grantor of Site[/SiteRef]
- Person is a recipient of Site[/SiteRef]
- Person is a previous grantor of Site[/SiteRef]
- Person is a previous recipient of Site[/SiteRef]
- Person is an occupant of a Site[/SiteRef]
- Person is a previous occupant of a Site[/SiteRef]
- Person is neighbour of Site[/SiteRef]
- Person is of Site[/SiteRef]

Clarifications:

- is tenant of – this is our default for a Person or Institution in possession of a location
- is landlord of – this should only be used in two circumstances: when a Person or Institution holds land from another Person or Institution, the latter is a landlord; when a Person or Institution receives rent from a location, they are a landlord.
- is occupant of – a Person is an occupant only if specified by the document and there is no evidence to suggest that he/she is a tenant (e.g. "the house occupied by John"; "the tenement where John dwells")
- is of – for use when the document gives the origins of a Person and said origins are not a Place (e.g. Thomas fitz Josce of Petergate, where Petergate is a street)

F) INSTITUTION-SITE[/SITEREF] ROLES:

- Institution is a landlord of Site[/SiteRef]
- Institution is a tenant of Site[/SiteRef]
- Institution is a previous landlord of Site[/SiteRef]
- Institution is a previous tenant of Site[/SiteRef]
- Institution is a grantor of Site[/SiteRef]
- Institution is a recipient of Site[/SiteRef]
- Institution is a previous grantor of Site[/SiteRef]
- Institution is a previous recipient of Site[/SiteRef]
- Institution is neighbour of Site[/SiteRef]

Clarifications:

- is tenant of – this is our default for a Person or Institution in possession of a location
- is landlord of – this should only be used in two circumstances: when a Person or Institution holds land from another Person or Institution, the latter is a landlord; when a Person or Institution receives rent from a location, they are a landlord.

G) ACTOR-SITE[/SITEREF] ROLES:

- Actor is a landlord of Site[/SiteRef]
- Actor is a tenant of Site[/SiteRef]
- Actor is a previous landlord of Site[/SiteRef]
- Actor is a previous tenant of Site[/SiteRef]
- Actor is a grantor of Site[/SiteRef]
- Actor is a recipient of Site[/SiteRef]
- Actor is a previous grantor of Site[/SiteRef]
- Actor is a previous recipient of Site[/SiteRef]
- Actor is neighbour of Site[/SiteRef]

Clarifications:

- is tenant of – this is our default for a Person or Institution in possession of a location
- is landlord of – this should only be used in two circumstances: when a Person or Institution holds land from another Person or Institution, the latter is a landlord; when a Person or Institution receives rent from a location, they are a landlord.

H) SITE[/SITEREF]-SITE[/SITEREF] ROLES:

- Site[/SiteRef] is part of Site[/SiteRef]
- Site[/SiteRef] is northern directional marker to Site[/SiteRef]
- Site[/SiteRef] is eastern directional marker to Site[/SiteRef]
- Site[/SiteRef] is western directional marker to Site[/SiteRef]
- Site[/SiteRef] is southern directional marker to Site[/SiteRef]
- Site[/SiteRef] is north-western directional marker to Site[/SiteRef]
- Site[/SiteRef] is north-eastern directional marker to Site[/SiteRef]

- Site[/SiteRef] is south-western directional marker to Site[/SiteRef]
- Site[/SiteRef] is south-eastern directional marker to Site[/SiteRef]
- Site[/SiteRef] is boundary to Site[/SiteRef]
- Site[/SiteRef] is breadth marker to Site[/SiteRef]
- Site[/SiteRef] is length marker to Site[/SiteRef]

Clarifications:

- Roads present a particular challenge. The typical example sees a road as a boundary to a location, with the road itself further specified as leading to x or leading from x to y. We have decided to connect x and y to the road using *is boundary to.*

I) SITE[/SITEREF]-PLACE[/PLACEREF] ROLES:

- Site[/SiteRef] is located in Place[/PlaceRef]

J) PERSON-PLACE[/PLACEREF] ROLES:

- Person is a landlord of Place[/PlaceRef]
- Person is a tenant of Place[/PlaceRef]
- Person is a previous landlord of Place[/PlaceRef]
- Person is a previous tenant of Place[/PlaceRef]
- Person is a grantor of Place[/PlaceRef]
- Person is a recipient of Place[/PlaceRef]
- Person is a previous grantor of Place[/PlaceRef]
- Person is a previous recipient of Place[/PlaceRef]
- Person is of Place[/PlaceRef]

Clarifications:

- is tenant of – this is our default for a Person or Institution in possession of a location
- is landlord of – this should only be used in two circumstances: when a Person or Institution holds land from another Person or Institution, the latter is a landlord; when a Person or Institution receives rent from a location, they are a landlord.
- is of – for use when the document gives the origins of a Person (e.g. Thomas fitz Josce of York)

K) INSTITUTION- PLACE[/PLACEREF] ROLES:

- Institution is a landlord of Place[/PlaceRef]
- Institution is a tenant of Place[/PlaceRef]
- Institution is a previous landlord of Place[/PlaceRef]
- Institution is a previous tenant of Place[/PlaceRef]
- Institution is a grantor of Place[/PlaceRef]
- Institution is a recipient of Place[/PlaceRef]
- Institution is a previous grantor of Place[/PlaceRef]

- Institution is a previous recipient of Place[/PlaceRef]
- Institution is located in Place[/PlaceRef]

Clarifications:

- is tenant of – this is our default for a Person or Institution in possession of a location
- is landlord of – this should only be used in two circumstances: when a Person or Institution holds land from another Person or Institution, the latter is a landlord; when a Person or Institution receives rent from a location, they are a landlord.
- is located in – for use when an Institution's location is specified (e.g. St Paul's in London)

L) ACTOR- PLACE[/PLACEREF] ROLES:

- Actor is a landlord of Place[/PlaceRef]
- Actor is a tenant of Place[/PlaceRef]
- Actor is a previous landlord of Place[/PlaceRef]
- Actor is a previous tenant of Place[/PlaceRef]
- Actor is a grantor of Place[/PlaceRef]
- Actor is a recipient of Place[/PlaceRef]
- Actor is a previous grantor of Place[/PlaceRef]
- Actor is a previous recipient of Place[/PlaceRef]

Clarifications:

- is tenant of – this is our default for a Person or Institution in possession of a location
- is landlord of – this should only be used in two circumstances: when a Person or Institution holds land from another Person or Institution, the latter is a landlord; when a Person or Institution receives rent from a location, they are a landlord.

M) PLACE[/PLACEREF]- PLACE[/PLACEREF] ROLES:

- Place[/PlaceRef] is part of Place[/PlaceRef]

N) PLACE[/PLACEREF]-SITE[/SITEREF] ROLES:

- Place[/PlaceRef] is boundary to Site[/SiteRef]

O) TRANSACTION ROLES:

P) PERSON-TRANSACTION ROLES:

- Person is grantor in Transaction
- Person is recipient in Transaction
- Person is previous grantor in Transaction
- Person is previous recipient in Transaction
- Person is a landlord in Transaction
- Person is a tenant in Transaction
- Person is a previous landlord in Transaction
- Person is a previous tenant in Transaction

- Person is participant in Transaction
- Person is previous participant in Transaction
- Person is witness to Transaction

Clarifications:

- is witness to – in a document which has two equal transactions (i.e. both in the document's present, and neither one a "main" transaction), you will have to link each witness to each transaction.

Q) INSTITUTION-TRANSACTION ROLES:

- Institution is grantor in Transaction
- Institution is recipient in Transaction
- Institution is previous grantor in Transaction
- Institution is previous recipient in Transaction
- Institution is a landlord in Transaction
- Institution is a tenant in Transaction
- Institution is a previous landlord in Transaction
- Institution is a previous tenant in Transaction
- Institution is participant in Transaction
- Institution is previous participant in Transaction

R) ACTOR-TRANSACTION ROLES:

- Actor is grantor in Transaction
- Actor is recipient in Transaction
- Actor is previous grantor in Transaction
- Actor is previous recipient in Transaction
- Actor is a landlord in Transaction
- Actor is a tenant in Transaction
- Actor is a previous landlord in Transaction
- Actor is a previous tenant in Transaction
- Actor is participant in Transaction
- Actor is previous participant in Transaction

S) SITE[/SITEREF]-TRANSACTION ROLES:

- Site[/SiteRef] is parcel in Transaction
- Site[/SiteRef] is location of Transaction [i.e. data/facta apud or dated at]

T) PLACE[/PLACEREF]-TRANSACTION ROLES:

- Place[/PlaceRef] is parcel in Transaction
- Place[/PlaceRef] is location of Transaction [i.e. data/facta apud or dated at]

U) TRANSACTION-TRANSACTION ROLES:

- Transaction precedes Transaction

To be considered:

- Transaction is part of Transaction

Clarifications:

- precedes – this carries forward, so if A<B<C, you do not also need to specify that A<C.

V) EVENT-TRANSACTION ROLES:

- Event precedes Transaction
- Event is concurrent with Transaction

W) EVENT-EVENT ROLES:

- Event precedes Event

X) DATE-DATE ROLES:

- Date precedes Date

Y) DOCUMENT-TRANSACTION ROLES:

- Document refers to Transaction

Z) DOCUMENT-DATE ROLES:

- Document is dated Date

AA) DOCUMENT-APPARATUS ROLES:

- Document contains Apparatus

V. APPENDIX TWO: TAXONOMY OF SITE ENTITIES

(Dr Stefania Merlo Perring, University of York)

Site entities are central to the objectives of ChartEx, concerning the relationships between people and ideas of space. The DM process required a database of 'Sites Types' with data captured from the collection of charters used for the ChartEx project. The entire collection of 'Charters of the Vicar Chorals' (Tringham 1993 and 2002) was selected because it constituted a large sample of 861 charter, concerning deeds in the city of York and in the North Yorkshire countryside . In contrast with other collection it had the advantage of providing a range of site types from both the urban (581 charters) and the rural (280 charters) contexts. Another advantage was that the edition of the charters of the Vicar Choral was in English, with a small proportion of charters in Latin; this was useful in the initial phase of the projects when the NLP component was not yet operative. The mixed language edition and the editorial notes to the English translation were also useful for recording some of the original site names in Latin, producing as much as possible a bilingual database. In the final part of the project the database of site names in Latin has been increased and enhanced with data harvested from 49 charters of the DEEDS database; however all these charters concern the rural context in Essex.

The DM process required that the database should be organised in taxonomy, to enable a short pathway for mining information. This was made by setting a typology, creating a hierarchy of categories with types and subtypes of sites, from the more general grouping down to the smallest details of sites. However, this is one of the many possible structures in which sites can be ordered and the resulting 'tree graph' reflects the conceptual categories that historians interested landscape and topography use to analyse space. It comprises of 5 Levels, each of which can be expanded with the addition of further categories and subcategories. This means that the taxonomy can be refined and increased with words and concepts from new documents from different regional and chronological contexts and in different languages.

An Excel database was designed to create the taxonomy and a diagram drawn in PowerPoint was prepared to visualise the 5 levels taxonomy as a tree-structure for the use of DM.

The Excel spreadsheets contained 154 entries and 6 columns, corresponding to the 5 levels of the taxonomy tree. Each level contained one or more categories (or classifications), which corresponded to branches in the tree-structure of the diagram. Level 5 (the more detailed containing the words used by documents describing in detail structures and natural features) comprises of two columns, one containing words in English and the second with the corresponding word in Latin. A further column with words in Old English or in the local idiom has been added.

Taxonomy of sites

Level 1

- Site

Level 2:

- Water
- Land.

Level 3:

- Water
  - water still
  - watercourse
- Land
  - agricultural
  - settlement
  - unknown

Level 4:

- Water
  - water still
  - watercourse
    - freshwater
    - drainage
- Land
  - agricultural
    - building
    - field
    - measure
    - natural
    - structure
    - tree
  - settlement
    - administrative
    - building
    - burial
    - dwelling

- infrastructure
- open area
- street system
- structure

- unknown

Level 5:

- Water
  - water still
    - pond, *stagnum*
  - watercourse
    - freshwater
      - river, *aqua*
      - stream, *rivosum*
      - water, *aqua*
    - drainage
      - channel, scorsum
      -
- Land
  - agricultural
    - building
      - windmill, *molendinum*
      - dovecote, *columbarium*
      - etc.
    - field
      - field, *campus*
      - meadow, *pratum*
    - measure
      - seld
      - selion
      - etc.
    - Natural
      - marshes, *mariscum*
      - etc.
    - structure
      - malt-kiln, *turallus*
      - shipfold, *ovile*
      - etc.
    - tree
      - ash, *fraxinum*
      - hedge
      - etc.
  - settlement
    - administrative
      - parish, *parochial*
      - etc.

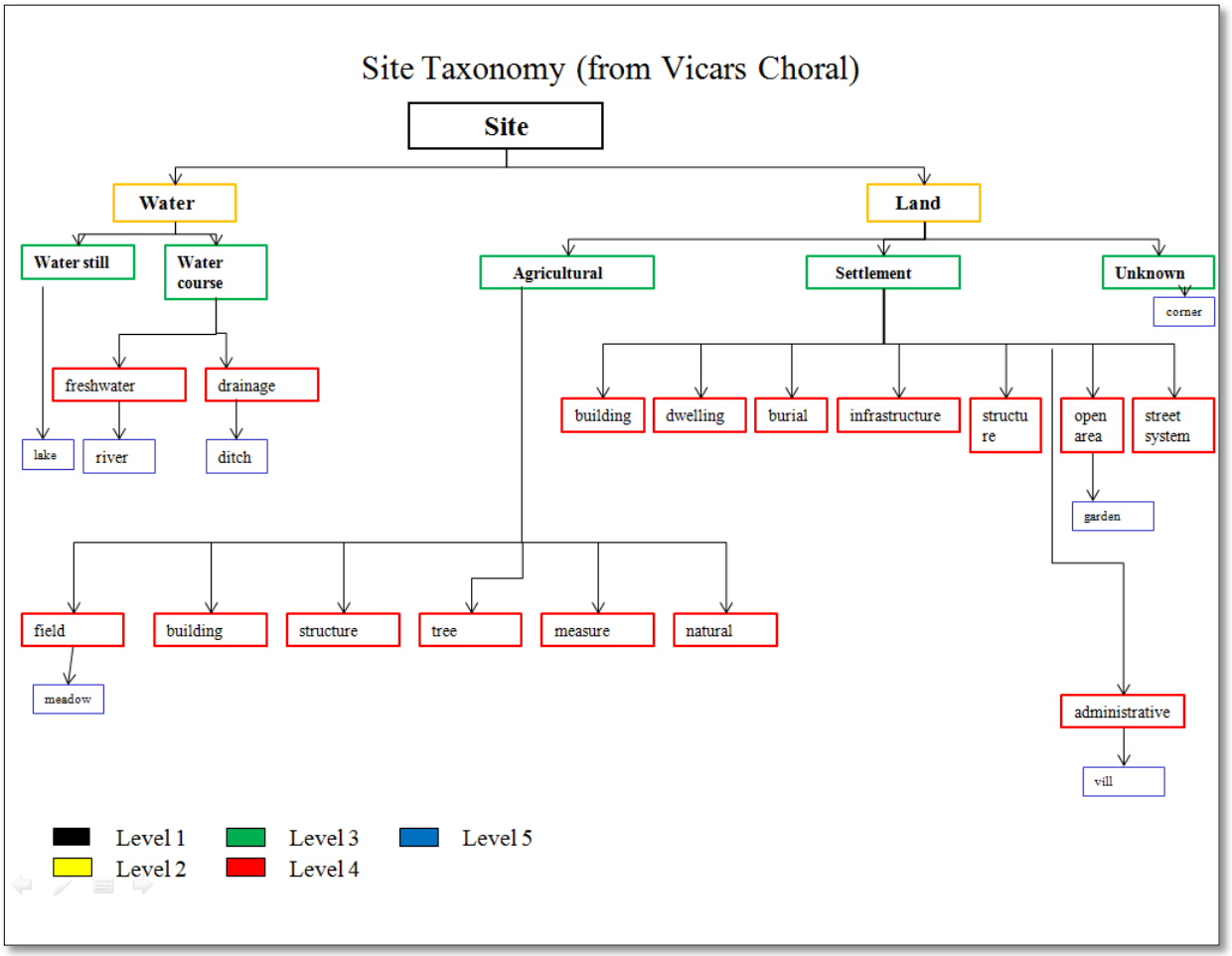
- Building
  - brew-house, *braseria*
  - castle, *castrum*
  - etc.
- burial
  - cemetery, *cimiterium*
  - etc.
- dwelling
  - messuage, *mesuagium*
  - vicarage, *vicaria*
  - etc.
- Infrastructure
  - market, *mercatus*, *forum*
  - bridge, *pons*
- open area
  - orchard, *ortu (hortus)*
  - etc.
- street system
  - street, *via, regia strata*
  - etc.
- Structure
  - wall, *murus*
  - fence, *palicium*
  - etc.

- Unknown
  - corner, *cornarium, angulus*
  - side, *latus*

- etc.

A. POSITIVE RESULTS OF INTERDISCIPLINARY COLLABORATION BETWEEN HISTORIANS AND DM COMPUTER SCIENTISTS

The collaboration between historians and DM scientists aimed at the creation of a tool, allowing a wider group of historians to explore the content of charters in a novel way, using the specialist skills of a small group of historians. These have collaborated with the DM scientists discussing research questions, explaining the reasoning underpinning their methods and checking and validating the scientists' work. DM has proved successful in replicating in a more rapid and efficient way previous labourious and time consuming methodologies, using a greater quantity of data. But the result of the collaboration has more profound implications for the historians, providing new perspectives in interpreting the content of charters, with the potential of innovate and opening new avenues in field of historical research.



These new perspectives are literally a physical point of view rather than and intellectual one, which is offered by the graphic display of networked entities produced by the DM process. For example the Biomine graph of the content of the 124 Vicar Choral charters used for this project represents a network of people, institutions, land, buildings, places events and archival collection numbers. This network connects human and non-human entities as expressed by the language of the charters, allowing previously unknown patterns to emerge generating new understanding. We may conduct a formal network analysis with the tools offered by the software, or be intrigued by a dense clusters of nodes with multiple ties. By analysing it we may discover that it represent land acquisition by one person or a conveyance of several properties to an institution. Or we may wish to investigate the meanings of a single node bridging together several other clusters, or ask why some small clusters have remained isolated. Another feature of the graph display of DM is that all entities are ordered in colour-coded groups and then listed in alphabetic order; from there they can be visualised in the network. This means that the content of large charter collections can be searched using new detailed indexes and catalogues.

The implications of DM graphs for historical research are that entities (people names, land, archival numbers etc.) and their relationships are freed from previous categories derived from the legal structure of the documents and by previous archival practices. Patterns observed in a network are starting points for analysis and can generate new research questions and insights. A network linking information from different archival fonds, such as charters, wills and cause papers, has the potential for creating alternative virtual archives with their own indexes. These archives can be created for a person or for a place. For example it would be possible to create personal archives for ordinary people of the past and to create archives relating to ordinary buildings; these will be useful to cultural historians interested in biographies of people and in material culture.

VI. APPENDIX III:

(Jon Crump, Robert Stacey, University of Washington)

The following text is extracted from Interim Narrative Summary for the Institute of Museum and Library Services Grant: LG-00-12-0455-12 submitted by Robert Stacey. It has been edited to remove repetition of material contained elsewhere in this white paper. It largely describes research into developing a LOD solution within ChartEx services which did not win the support of all partners. Nevertheless its worth as an experimental approach to the data is indicated by the success of other recently funded projects such as AHRC Big Data SNAP:DRGN http://www.kcl.ac.uk/artshums/depts/ddh/newsrecords/2014/snapdrgn.aspx.

A. SUPPORTING THE BRAT ANNOTATIONS

Washington took on the task of installing and maintaining the BRAT tool for use by the CHARTEX project for marking up our test data.

One of the essential challenges of the CHARTEX project was to see if we could usefully process charter data coming from extremely heterogeneous sources. This included everything from scanned published documents, and digitized full-text charters in Latin scraped from web-sites or derived from data-bases; to English summaries of charters provided to us as spreadsheets, and as XML files in several different formats. In order to prepare this wide range of sources for annotation, we developed and have maintained a set of ad hoc computer scripts written in the Python programming language to extract the text data from our chosen corpora and generate the plain-text files needed by the annotation tool.

B. LINKED OPEN DATA

LOD was not part of the project proposal. Nevertheless early in the project we advocated and it was agreed that our data should take the form of RDF triples (subject, predicate, object statements about resources), and that those triples should be ordered by our own "Place and Site" oriented vocabulary. That vocabulary in turn would sub-class the CIDOC CRM ontology in order to make our data interoperable with other systems and publishable as "Linked Open Data." RDF, (the Resource Description Framework) is the lingua franca that allows the semantic web to work. Along with OWL (Web Ontology Language), RDF makes possible a web of information, not merely of documents. These conventions make it possible for researchers to add to and comment upon the available information in a given domain in a distributed and decentralized way using generalizable "upper ontologies" like the CIDOC CRM. (Comite International pour la Documentation, Conceptual Reference Model) The CIDOC CRM has emerged in the past few years as the most robust effort to provide a context for resource description and interoperability in various cultural heritage domains: specifically in museum studies, but also notably in archaeology. It has obvious applicability in the domain of history as well but is not generally used by archive services. We entered into our LOD research in an experimental spirit.

After the partners had agreed upon version 1 of our ontology, we codified those guidelines as a formal configuration document for the BRAT annotation tool. York then codified the guidelines as RDF/OWL and we collaborated to refine and correct that provisional ontology. Given the nature of the data as RDF triples, there was a range of preliminary experiments and development activities that suggested themselves, and that were possible using available resources: developing mechanisms for harvesting the data created by the annotation process; generating RDF graphs from those data; storing, retrieving, and querying those graphs; and displaying the results of those operations usefully in a web page interface.

For the purpose of those small scale experiments, we used the in-memory triple store afforded by the Python library RDFlib. This library also implements the SPARQL 1.1 specification for querying RDF triples and a range of other well-developed functionality for generating and manipulating RDF graphs. The in-memory store could serve as the basis for

experimentation, and the other functionality of RDFlib would be developed to provide the server-side services for the workbench.

Throughout the fall 2012 we made progress developing functions to parse the output of the BRAT annotation tool and generate RDF triples from those data. We developed a mechanism based on the “dot” language and the “graphviz” graph visualization library for visualizing RDF graphs and for displaying the result in a web page using SVG. It is worth noting in passing that in addition to preparing the way for visualizations for the workbench, the graphviz visualizations proved immediately useful as an error checking mechanism for the ongoing test-data markup of the Cluny charter material that was still being conducted by Columbia.

We also developed functions to query the RDF graphs through a web interface by means of the SPARQL query language, and by plain-text searches using the unix GREP utility, and “fuzzy” searching using a Python implementation of the Damerau-Levenshtein distance measurement algorithm. Additionally we experimented with an interface for enriching our data by accessing the APIs of other information services like google books.

These developments at first were confined to document level data. In the winter we began to grapple with the problems associated with generating and manipulating graphs of corpora of documents. We created a user interface for generating RDF graphs of the data in whole directories of our BRAT-annotated test data.

To this point we had been using URIs for our data coined on an ad-hoc basis. For the Linked Open Data paradigm to work, URIs must be constructed on a rational basis that provides for orderly dereferencing and redirection mechanisms. In re-assessing our progress to date, we observed that the base data we had been using, the BRAT annotation files, were actually most accurately described as collections of blank nodes. We developed parallel functions to model our data to reflect this observation. This effort confirmed the need to name our entities and relations in order to support more transparent queries, and, more important, to support the annotation facilities we envisioned for the workbench.

After our meeting in York in January2013, York and Brighton established server space for the CGI workbench web application, and arranged for the use of an enterprise level triple store provided by York’s Archaeological Data Service. This triple store, implemented as a Franz AllegroGraph server, we have access to only through its REST/HTTP interface.

Since those meetings we have made rapid progress in adapting our earlier experiments to the environment of what will be our production triple store. Using Python libraries for REST interactions, we have developed web interfaces for uploading, retrieving, and querying our data. We also made progress toward providing user annotation functionality.

Since 2009, most triple-stores, including AllegroGraph, have provided “namedgraph” functionality that goes beyond the basic RDF specification. Also since 2009, extensive work has been done by the W3C Open Annotation Community Group on creating a standardized ontology for the creation of RDF annotations. The Open Annotation Data Model will be rolled out next month at Stanford University, and in June at the University of Manchester in the UK. We are now working to implement this Open Annotation ontology to make it possible for our

users to create annotations for any entity or relation in our data, and to store those annotations on the same server and in the same form as the target data. This will make it possible for users to assess the results of the NLP and DM projects, and to collaboratively publish their findings as linked open data that can be queried, deployed, and re-purposed as needed along with our original charters that have been marked-up either manually, or by means of NLP processing.