Managing Meta Data in a Research Collection of Medieval Latin Charters

Michael Gervers & Michael Margolin

Introduction

The DEEDS Project was founded at the University of Toronto in 1975 to provide computerized access to the content of twelfth- and thirteenth-century English conveyances, the great majority of which were issued without dates, making it extremely difficult to determine the nature of economic, social and administrative change other than in very general chronological terms. Members of the DEEDS team, and especially my colleague, Michael Margolin, have long worked to develop a program to assist in establishing dates for that undated majority. To that end, and with the assistance of numerous graduate students, we have developed a database of over 9,500 charters from our period, dated either internally, or by the editor, to within a year of the actual date of issue. To do this we have searched through over 100,000 documents in 190 printed sources. Charter metadata includes chronological and topographical information, details of the religious house or noble household in whose archive the document has been preserved, and the roles, titles and social status of the individuals involved. Aspects of the document itself are also covered by the metadata: its type, its legal purpose, whether it is an original or a copy, and its archival status. Moreover it is here that diplomatic encoding of the charter is carried out. Names of persons and places are encoded separately. In order to avoid overlap, a layering system is used.

The purpose of diplomatic encoding for us is to provide the means to identify degrees of change within the parts of the charter. Little is known about the “why” and “when” of legal developments, although it has been shown that small differences in wording reflect changes in terms of the historical context,[1] and the need for clarification and precision to matters concerning the transfer of property.

There is also the all-important question of authenticity. We know that many forgeries do exist, but it can be very difficult to distinguish between those documents which are entirely authentic and those to which changes have been made, especially in later copies. Diplomatic encoding can be a useful means of identifying words, phrases and grammatical constructions which appear inappropriate for the purported date of writing. There is much evidence that later changes to the content of medieval conveyances did not necessarily constitute what we today think of as forgery. The scribe or notary may just have been trying to bring a document up to date; that is, to reflect the contemporary situation. Following the parchment trail in search of such changes is invariably restricted by the incomplete survival of the historical record, but the use of sophisticated statistical technology can help to fill information gaps.

Encoding DEEDS Medieval Latin Charters

In the early days of the DEEDS Project we used a relational database management system (DBMS) to support academic research. The text of the charter was coded in flat files while some metadata was stored in DBMS table fields. Links between charter text and metadata were implemented by DBMS constraints.

By the mid-nineties the number of charters had grown significantly, the spectrum of research had widened and a total redesign of the core architecture of the system became necessary. Extensible Markup Language (XML) was the obvious choice because of its ability to reproduce the structural hierarchy of a DBMS. XML is derived from Standard Generalized Markup Language (SGML), but is an ISO standard in its own right. It has a set of rules for describing data and designing text formats that allows people to structure their own data. It uses a Document Type Definition (DTD) or an XML Schema to describe the data. At that time encoding rules were governed by the DEEDS proprietary research-oriented DTD. The transition from DBMS to XML-based encoding allowed greater flexibility in how metadata was structured and linked to the text of the charter. In particular, the diplomatic parts of the charter could easily be added to the metadata of the charter.

By the end of the 20th century a number of internationally recognized standards for electronic text encoding had emerged and become established. The most widely known and used standards for the digitization of texts were the Text Encoding Initiative (TEI), Encoded Archival Description (EAD) and Dublin Core. These, along with many others, are entirely based on XML/SGML, a well known but somewhat rusty encoding concept based initially exclusively on DTD, but later enhanced by XML Schema support.

Recent years have seen emerging principles of Semantic Web and new information encoding standards such as Resource Description Framework (RDF) and Online Web Language (OWL) which laid the foundation of a new decentralized platform for distributed knowledge. It was originally created in 1999 as a standard on top of XML for encoding metadata and had evolved to the 3WC standard by 2004. OWL is the next evolutionary step in the implementation of Semantic Web principles from RDF which provides a generic, flexible way to break the piece down into small parts, called triples. Initially the information is broken down into a labeled, directed graph. Each edge of the graph represents a fact or a relationship between two things, divided into three parts: subject, predicate, and object, that is, what is at the start of the edge, the type of edge, and what is at the end of the edge.

RDF is serialized using XML but the fundamental model is completely independent of XML. An XML binds a strict hierarchical structure of metadata to a text. RDF deals with a generic description of the relationships between resources. RDF incorporates features to facilitate data merging even if the underlying Schemas differ, and it specifically supports the evolution of Schemas over time without making changes to the data. In general, RDF offers more robust protection for Schemas and Ontologies than XML does for Schemas. RDF data can routinely refer to several different terminologies (in the form of RDF Schemas or Ontologies) developed by various communities. Mixing those terminologies within the same RDF dataset is easy and natural, whereas mixing different XML Schemas within the same XML data set, in spite of the availability of namespaces, remains rather complicated.

The conversion of the entire DEEDS collection of charters from plain XML to RDF was a logical evolutionary step towards further enhancement of its research capabilities and manageability. This conversion was motivated by a desire to increase the flexibility of encoding and to provide a generic interface for the encoded information. DEEDS charters are now encoded using so-called "Stripped Syntax", a convention which allows RDF to be expressed as XML. Current charter encoding incorporates many RDF elements defined by RDF vocabulary description language, RDF Schema (RDFS). Examples of charter encoding before (Fig. 1) and after (Fig. 2) conversion from XML to RDF are shown here.

Along with the intrinsic flexibility of RDF encoding, the present implementation of the DEEDS charter repository also takes full advantage of the latest developments in current DBMS support for XML. Each charter is encoded as an RDF/XML document and is later placed into a native database XML type of storage. XML type natively supports the use of XQuery and XPath statements embedded into Structured Query Language (SQL) queries. Additionally, a DBMS provides facilities for the creation of database indexes for all XML elements of an underlying RDF/XML document. When indexing is completed any encoded information can be retrieved using the DBMS native low level high speed interface.

This DBMS-based organization of charter repositories provides enough convenience and performance for routine archival activities like searching and browsing and is well suited to hosting a source charter collection. However when the researcher needs to draw inferences about charter properties from statistical computations the DBMS organization is barely capable of delivering adequate application performance. The DEEDS Project program for chronological charter attribution, for example, must execute hundreds of thousands of queries to generate a final result. While some query execution performance could be improved by using a native database interface as an alternative to SQL, other issues such as implicit DBMS overhead are virtually impossible to overcome. The DEEDS solution for that problem was the creation of a two-tier charter repository organization. The first tier consists of RDF/XML encoded charter documents which are stored in DBMS. This tier provides Internet exposure for the charter repository accompanied by search and browse features. The second tier is composed of sets of charter documents, generated on demand, which carry a limited set of problem-oriented metadata. Each such set is later indexed by the text search engine which builds in memory indexes while at the same time allowing a low level programming interface of metadata and content.

Metadata

Metadata encoded into DEEDS charters is by nature more research oriented than archival. It is gathered from differing sources such as the editorial notes supplied with a printed source, knowledge of the charter's external source, analytical information compiled by studying the charter text and information about charter diplomatics.

All the first tier metadata is encoded in RDF/XML and can be divided into two major groups: one directly linked to the text of the charter and another, not directly associated with any particular part of the text, which refers to the whole charter. Examples of encoding of the "name" node in XML (Fig. 3) and later in RDF/XML (Fig. 4) are shown here. A metadata group which maintains links to the text relies on RDF and XML elements defined according to the various parts of the charter text. A distinctive feature of that group is an implementation of encoding.

The DEEDS Project has developed a unique approach to practical encoding by mapping metadata to a particular area of the text instead of embedding tags into it. According to this model, encoding is shaped as a collection of “layered” elements. Each element is tightly linked to a specific category. Physical tags can be inserted later dynamically, on demand, by using mapping addresses stored in "layer" elements. There is no limitation to the number or order of "layer" elements. They are absolutely independent and can overlap to a degree that is not allowed under conventional XML encoding. This encoding model can also take full advantage of sophisticated client interfaces provided by so-called Rich Internet Application (RIA) frameworks (like Adobe "Flex") which have data service entirely separated from presentation objects.

Another metadata group is encapsulated inside an RDF "descriptive" element, attached to the root of the document hierarchy. RDF containers are resources used to represent collections which are frequently used for encoding of metadata. RDF containers like "bag" ("Bag") and "seq" ("Sequence") have proved to be particularly useful in charter encoding.

In the second tier of charter repository organization, document encoding is implemented in HTML and metadata is encoded using "meta" tags. Both document and metadata are generated from a corresponding RDF/XML source by an in-house developed program and then stored in the file system. Later "meta" tags are indexed and cashed by an indexing engine. Each of the second tier set of documents can be regenerated or updated by using the same program. It is obvious that maintenance of a two-tier system requires additional effort and storage space but improved performance and enhanced flexibility make such a system a viable option in the research environment.

The first tier of the DEEDS charter collection is maintained exclusively through a program called "Document Manager" (DM), which has been developed by the DEEDS Project. DM has a built-in intuitive graphical user interface which shields the end user completely from the complexities of underlying RDF/XML encoding. DM provides validation and generates encoding from information entered into various text fields. The second tier is maintained by the use of a set of command line utilities executed in batch.

Attribution of Medieval Latin Charters

Attributing and particularly establishing a date for a charter has always been a main goal of the DEEDS Research Project. Since approximately 92% of the estimated one million or more English charters written in Latin which survive from the 12th and 13th centuries as originals or copies are without dates, the development of accurate methods for chronological evaluation are very important. Along with chronological attribution there is also a need for scribal identity verification which could possibly even indicate forgeries. The DEEDS Project has developed two online programs for evaluation of chronology and scribal identity which are currently at the beta stage.

The **Chronology Evaluation program** uses vocabulary analyses and a statistical engine to produce a final result. To establish a chronology the program extracts vocabulary patterns from the given text and searches for identical patterns among all securely dated charters. It then records the date of charters in which the same pattern occurs. Information gathered from these patterns is then passed to the statistical engine, which splits it into several independent data flows. Each data flow is then processed separately. The results of this execution are later presented by graphical output and statistical estimate.

For the further reduction of errors and the resolution of cases where the default execution path has produced an ambiguous result, a number of alternative path options are available. For example the total number of documents against which vocabulary patterns are to be matched can be reduced, by subdivision according to type of charter content (“grant”, “lease”, “agreement”, etc), or charter geography, or chronological span. There are also a number of lexical transformations which could be applied to the text which is being examined, in any combination with scope reduction options. Vocabulary patterns are always matched to a corresponding second tier set of charter documents. Lexical transformations can include spelling normalization and full or partial lemmatization. An example of program output using a default execution procedure is illustrated by Fig 5.

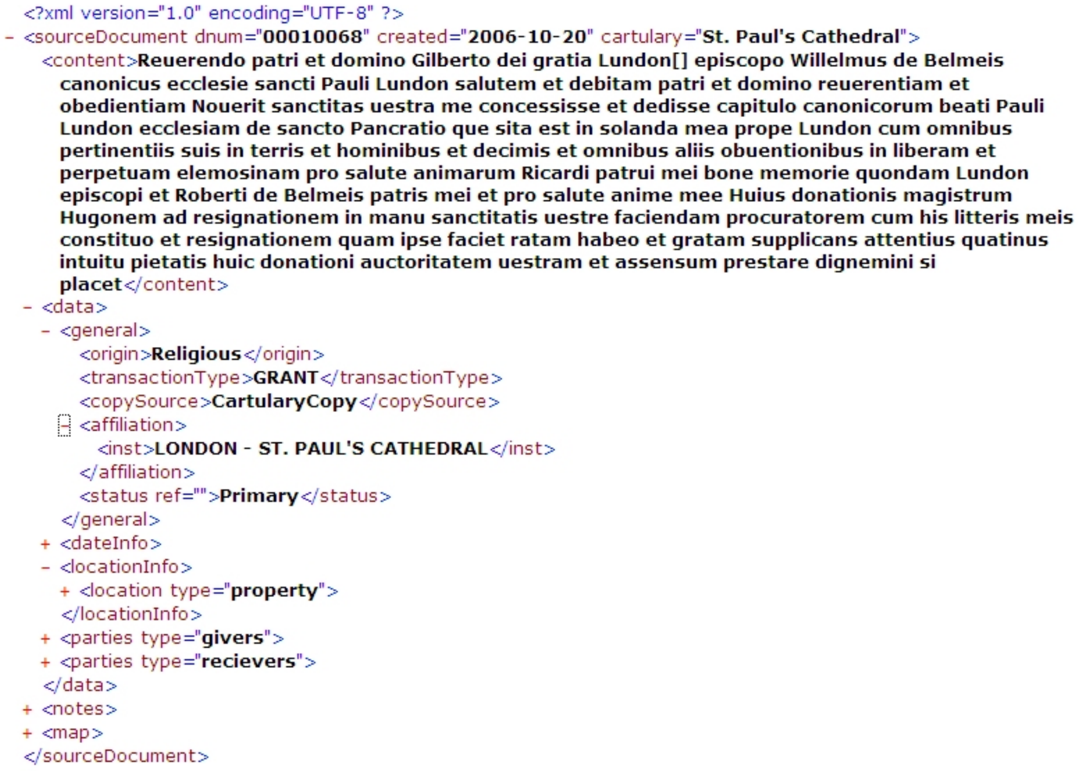
An algorithm prepared with the assistance of the Department of Statistics at the University of Toronto using a comparison of word-order between dated and undated charters is an another way of establishing chronological boundaries for any given charter.

The **Scribal Identity Verification program** employs two stylometric algorithms: Cumulative Sum Charts (QSUM) and Entropy for Markov Chains of Letters (EMCL). This variation of QSUM technology is fairly well known and has been widely adapted for authorship identification. It is plotted graphically, based on a sequence of measures calculated by using the deviation of stylometric measures in individual sentences. The principal impediment to the application of the QSUM technology to medieval Latin charters was an insufficient number of unambiguously identifiable and complete sentences in the charter text. A solution to the problem was found by substituting artificially emulated groups of sequential words for real sentences. These groups are extracted from the text by using an algorithm similar to the one which the chronology attribution program employs to extract vocabulary patterns. The resultant distribution of stylometric measures is then passed to the statistical engine which delivers a final QSUM chart for each document. The EMCL technique is based on the Markov model for sequence of letters in the text. It is assumed that the probabilities of transition between pairs of letters are specific to the author of the text. The program generates letter to letter transition matrixes for the text being examined and later compares them. A sample of output from the Scribal Identity program is shown in Fig. 6.

In Closing

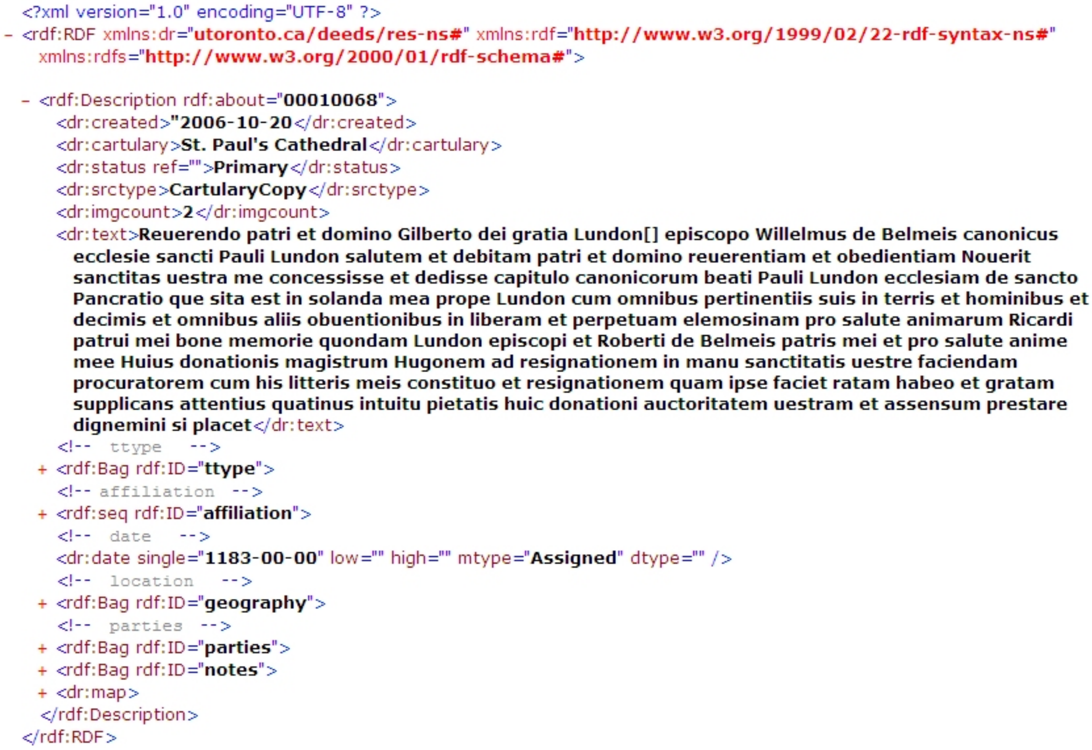
Charters from the 12th and 13th century are a significant source for the study of English medieval history of the period. The application of modern technologies to the interpretation of the charter, coupled with the enormous and ever-increasing power of the computer, allows us to uncover new and heretofore virtually inaccessible pieces of information about social, economic and administrative change at the time. The searchable database of dated charters from the period has provided the DEEDS Research team with a corpus of material on which the innovative dating and authorship identification programs are based. This will be of primary importance in establishing a more accurate chronology for the charters which have already been published, and for the hundreds of thousands which remain as yet unread and unedited.

Appendix 1



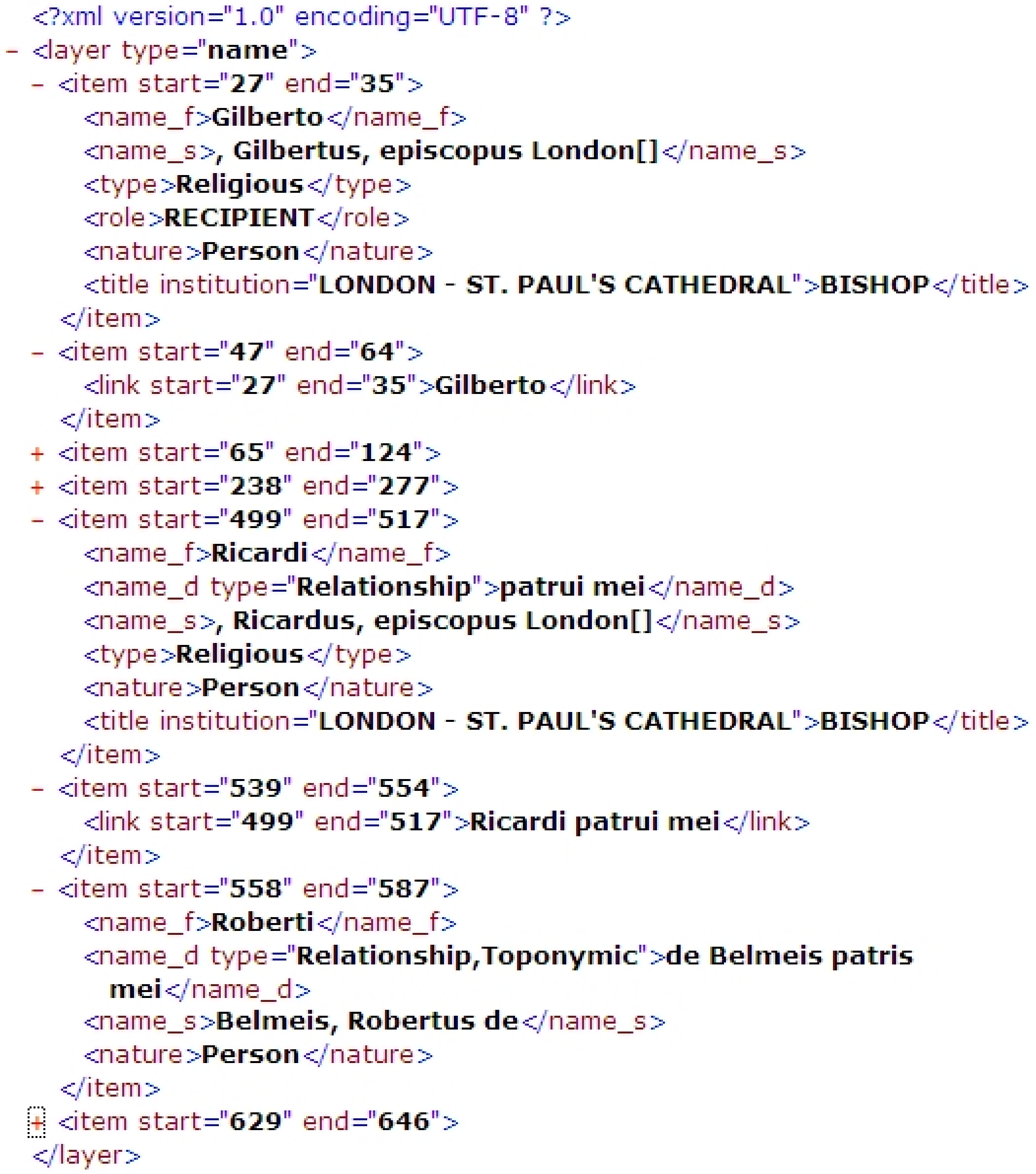
```
  <?xml version="1.0" encoding="UTF-8" ?>
- <sourceDocument dnum="00010068" created="2006-10-20" cartulary="St. Paul's Cathedral">
    <content>Reuerendo patri et domino Gilberto dei gratia Lundon[] episcopo Willelmus de Belmeis
      canonicus ecclesie sancti Pauli Lundon salutem et debitam patri et domino reuerentiam et
      obedientiam Nouerit sanctitas uestra me concessisse et dedisse capitulo canonicorum beati Pauli
      Lundon ecclesiam de sancto Pancratio que sita est in solanda mea prope Lundon cum omnibus
      pertinentiis suis in terris et hominibus et decimis et omnibus aliis obuentionibus in liberam et
      perpetuam elemosinam pro salute animarum Ricardi patrui mei bone memorie quondam Lundon
      episcopi et Roberti de Belmeis patris mei et pro salute anime mee Huius donationis magistrum
      Hugonem ad resignationem in manu sanctitatis uestre faciendam procuratorem cum his litteris meis
      constituo et resignationem quam ipse faciet ratam habeo et gratam supplicans attentius quatinus
      intuitu pietatis huic donationi auctoritatem uestram et assensum prestare dignemini si
      placet</content>
  - <data>
    - <general>
        <origin>Religious</origin>
        <transactionType>GRANT</transactionType>
        <copySource>CartularyCopy</copySource>
      - <affiliation>
          <inst>LONDON - ST. PAUL'S CATHEDRAL</inst>
        </affiliation>
        <status ref="">Primary</status>
      </general>
    + <dateInfo>
    - <locationInfo>
      + <location type="property">
      </locationInfo>
    + <parties type="givers">
    + <parties type="recievers">
    </data>
  + <notes>
  + <map>
  </sourceDocument>
```

Figure 1 St. Paul's Cathedral Cartulary, Charter #68 using XML



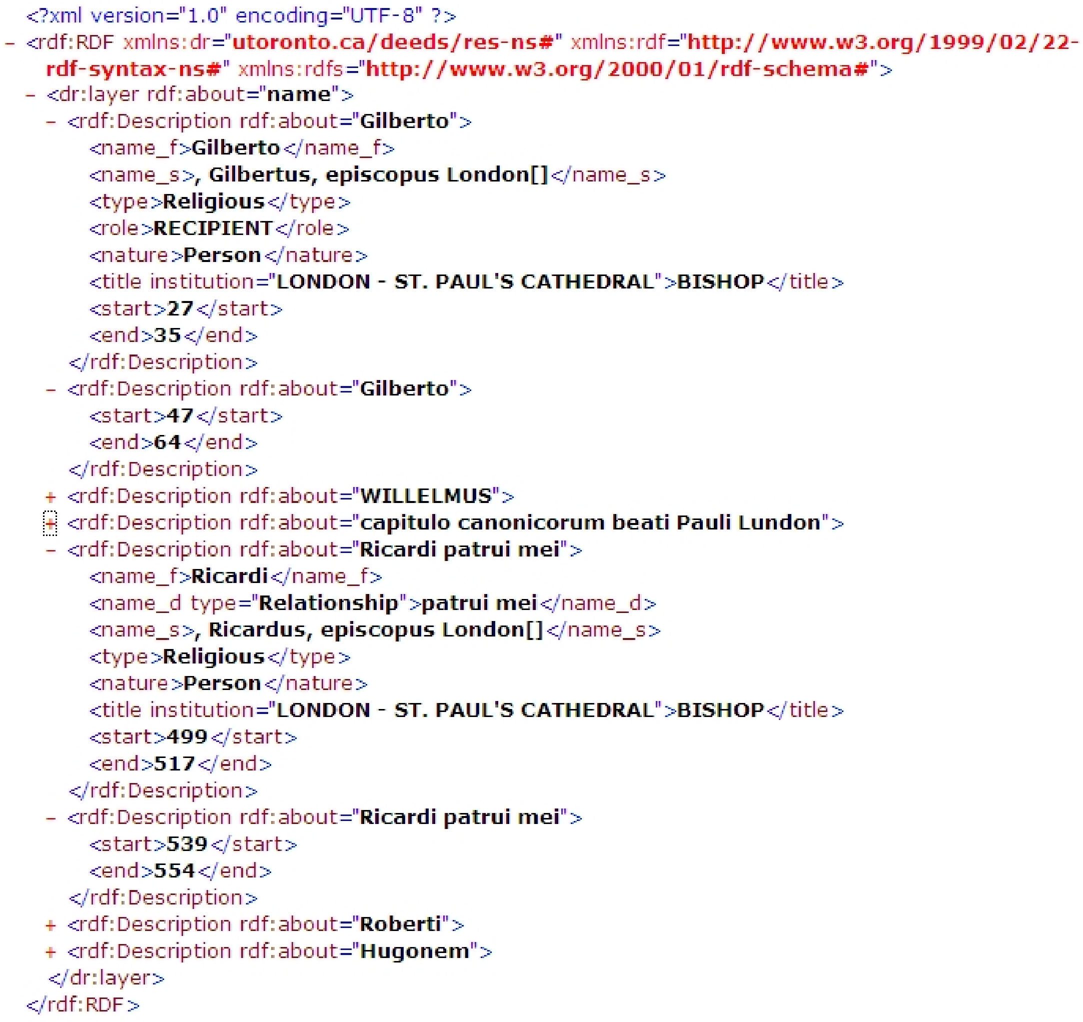
```
  <?xml version="1.0" encoding="UTF-8" ?>
- <rdf:RDF xmlns:dr="utoronto.ca/deeds/res-ns#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

  - <rdf:Description rdf:about="00010068">
      <dr:created>"2006-10-20</dr:created>
      <dr:cartulary>St. Paul's Cathedral</dr:cartulary>
      <dr:status ref="">Primary</dr:status>
      <dr:srctype>CartularyCopy</dr:srctype>
      <dr:imgcount>2</dr:imgcount>
      <dr:text>Reuerendo patri et domino Gilberto dei gratia Lundon[] episcopo Willelmus de Belmeis canonicus
        ecclesie sancti Pauli Lundon salutem et debitam patri et domino reuerentiam et obedientiam Nouerit
        sanctitas uestra me concessisse et dedisse capitulo canonicorum beati Pauli Lundon ecclesiam de sancto
        Pancratio que sita est in solanda mea prope Lundon cum omnibus pertinentiis suis in terris et hominibus et
        decimis et omnibus aliis obuentionibus in liberam et perpetuam elemosinam pro salute animarum Ricardi
        patrui mei bone memorie quondam Lundon episcopi et Roberti de Belmeis patris mei et pro salute anime
        mee Huius donationis magistrum Hugonem ad resignationem in manu sanctitatis uestre faciendam
        procuratorem cum his litteris meis constituo et resignationem quam ipse faciet ratam habeo et gratam
        supplicans attentius quatinus intuitu pietatis huic donationi auctoritatem uestram et assensum prestare
        dignemini si placet</dr:text>
      <!--  ttype   -->
    + <rdf:Bag rdf:ID="ttype">
      <!-- affiliation  -->
    + <rdf:seq rdf:ID="affiliation">
      <!--  date   -->
      <dr:date single="1183-00-00" low="" high="" mtype="Assigned" dtype="" />
      <!--  location   -->
    + <rdf:Bag rdf:ID="geography">
      <!--  parties  -->
    + <rdf:Bag rdf:ID="parties">
    + <rdf:Bag rdf:ID="notes">
    + <dr:map>
    </rdf:Description>
  </rdf:RDF>
```

Figure 2 St. Paul’s Cathedral Cartulary, Charter #68 using RDF/XML



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <layer type="name">
  - <item start="27" end="35">
      <name_f>Gilberto</name_f>
      <name_s>, Gilbertus, episcopus London[]</name_s>
      <type>Religious</type>
      <role>RECIPIENT</role>
      <nature>Person</nature>
      <title institution="LONDON - ST. PAUL'S CATHEDRAL">BISHOP</title>
    </item>
  - <item start="47" end="64">
      <link start="27" end="35">Gilberto</link>
    </item>
  + <item start="65" end="124">
  + <item start="238" end="277">
  - <item start="499" end="517">
      <name_f>Ricardi</name_f>
      <name_d type="Relationship">patrui mei</name_d>
      <name_s>, Ricardus, episcopus London[]</name_s>
      <type>Religious</type>
      <nature>Person</nature>
      <title institution="LONDON - ST. PAUL'S CATHEDRAL">BISHOP</title>
    </item>
  - <item start="539" end="554">
      <link start="499" end="517">Ricardi patrui mei</link>
    </item>
  - <item start="558" end="587">
      <name_f>Roberti</name_f>
      <name_d type="Relationship,Toponymic">de Belmeis patris
        mei</name_d>
      <name_s>Belmeis, Robertus de</name_s>
      <nature>Person</nature>
    </item>
  + <item start="629" end="646">
  </layer>
```

Figure 3 "Name" layer encoding using XML



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <rdf:RDF xmlns:dr="utoronto.ca/deeds/res-ns#" xmlns:rdf="http://www.w3.org/1999/02/22-
    rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  - <dr:layer rdf:about="name">
    - <rdf:Description rdf:about="Gilberto">
        <name_f>Gilberto</name_f>
        <name_s>, Gilbertus, episcopus London[]</name_s>
        <type>Religious</type>
        <role>RECIPIENT</role>
        <nature>Person</nature>
        <title institution="LONDON - ST. PAUL'S CATHEDRAL">BISHOP</title>
        <start>27</start>
        <end>35</end>
      </rdf:Description>
    - <rdf:Description rdf:about="Gilberto">
        <start>47</start>
        <end>64</end>
      </rdf:Description>
    + <rdf:Description rdf:about="WILLELMUS">
    + <rdf:Description rdf:about="capitulo canonicorum beati Pauli Lundon">
    - <rdf:Description rdf:about="Ricardi patrui mei">
        <name_f>Ricardi</name_f>
        <name_d type="Relationship">patrui mei</name_d>
        <name_s>, Ricardus, episcopus London[]</name_s>
        <type>Religious</type>
        <nature>Person</nature>
        <title institution="LONDON - ST. PAUL'S CATHEDRAL">BISHOP</title>
        <start>499</start>
        <end>517</end>
      </rdf:Description>
    - <rdf:Description rdf:about="Ricardi patrui mei">
        <start>539</start>
        <end>554</end>
      </rdf:Description>
    + <rdf:Description rdf:about="Roberti">
    + <rdf:Description rdf:about="Hugonem">
    </dr:layer>
  </rdf:RDF>
```

Figure 4 "Name" layer encoding using RDF/XML

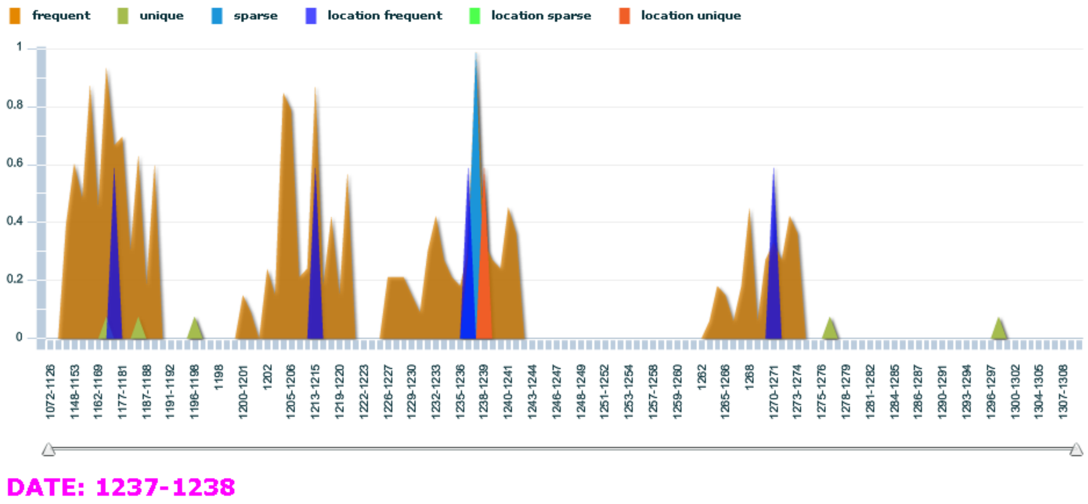


Figure 5 Charter # 94, Colchester- St. John the Baptist. December 1237

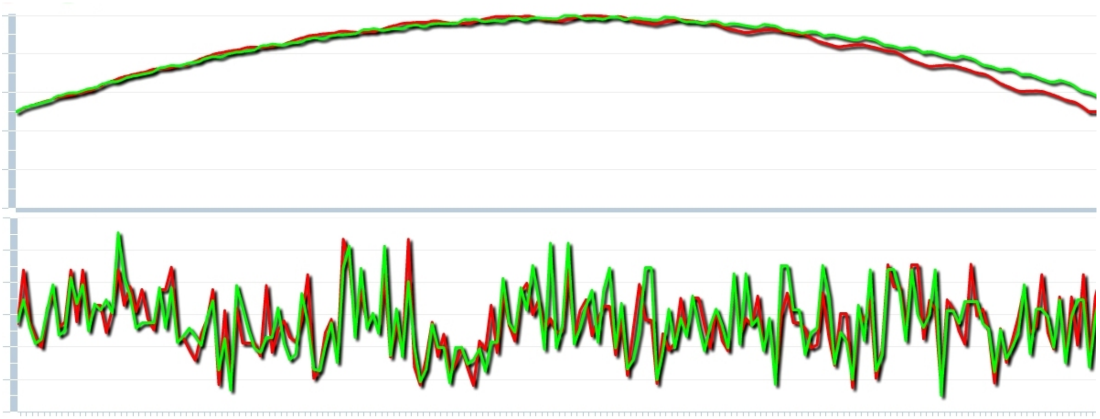


Figure 6 QSUM and EMCL results for William son of Derkin, British Library Cotton Ms Nero E VI, Hospitaller Cartulary Secunda Camera nos. 462 and 512

Bibliography

1. D.V. Khmelev, *Disputed Authorship Resolution through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language* Texts Journal of Quantitative Linguistics 2000, Vol. 7, No. 3, pp. 201-207
2. *Extensible Markup Language (XML) 1.0*, Second Edition, T. Bray, J. Paoli, C.M. Sperberg-McQueen and E. Maler, Editors. World Wide Web Consortium. 6 October 2000.
3. Farringdon, Jill, *Analysing for Authorship: A Guide to the Cusum Technique*, Cardiff: University of Wales
4. Press, 1996.
5. Holmes, D. I., *Authorship Attribution*, Computers and the Humanities, 1994.
6. Michael Gervers, ed. *The Cartulary of the Knights of St. John of Jerusalem in England, Part 1, Secunda Camera*. London: The British Academy, 1982
7. Michael Gervers, Michael Margolin, *New Methods For The Analysis Of Digitized Medieval Latin Charters*. Conference paper, Berlin 2006. http://res.deeds.utoronto.ca:49838/pubs/
8. Michael Gervers, Michael Margolin, *Towards The Content Analyses Of English Private Charters Of The Twelfth And Thirteenth Centuries*. Conference paper, Munich 2004. http://res.deeds.utoronto.ca:49838/pubs/
9. OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-guide/

10. RDF/XML Syntax Specification (Revised) W3C Recommendation 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/
11. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-schema/ .
12. TEI (The Text Encoding Initiative). http://www.tei-c.org/
13. The DEEDS Database of Mediaeval Charters: Design and Coding for the RDBMS Oracle 5 Michael Gervers, Gillian Long, and Michael McCulloch History & Computing Vol 2 No 1 1990. Oxford University Press, Pinkhill House, Southfield Road, Eynsham, Oxford, OX8 1JJ.
14. G.Tummarello, C.Morbidoni, E. Pierazzo, "*Toward textual encoding based on RDF*", ICCC 9th International Conference on Electronic Publishing, June 2005, Leuven, Belgium.
15. XML Path Language (XPath). Version 1.0. W3C Recommendation 16 November 1999. http://www.w3.org/TR/xpath
16. XML Path Language (XPath) 2.0. W3C Recommendation 23 January 2007. http://www.w3.org/TR/xpath20/

[1] Michael Gervers, “Changing Forms of Hospitaller Address in English Private Charters of the Twelfth and Thirteenth Centuries”, in ***The Crusades and the Military Orders: Expanding the Frontiers of Medieval Latin Christianity***, ed. Zsolt Hunyadi & József Laszlovszky, Budapest: Central European University, 2001 (ISBN 963-9241-42-3), 395-405.