INTERNATIONAL CONFERENCE

DIGITAL DIPLOMATICS. TOOL FOR THE DIGITAL DIPLOMATIST

Università degli Studi di Napoli Fedrico II, Palazzo degli Uffici, 29th and 30th September 2011
Società Napoletana di Storia Patria, 1st October 2011

BOOKLET

Presentation

Scholars of diplomatics never had a fundamental opposition on using modern technology to support their research. Nevertheless no technology since the introduction of photography had such an impact on questions and methods of diplomatics as the computer had: Digital imaging gives us cheap reproductions at high quality, so nowadays large copora of documents are to be found online. Digital imaging allows manipulations to make apparently invisible traces visible. Modern information technology gives us access to huge text corpora in which single words and phrases can be found thus helping to indicate relationsships, to retrieve parallel texts for comparision or plot geographical and temporal distributions.

The conference aims at presenting projects which working to enlarge the digitised charter corpus on the one hand and on the other hand will put a particular focus on research applying information technology on medieval and early modern charters aiming at pure diplomatic questions as well as historic or philologic research.

The organizer of the conference therefore have invited papers dealing with questions like:

- How can we improve the access to digital charter corpora?
- How can the presentation of digital charter corpora help research with them?
- Are there experiences in the application of complex information technologies (like named entity recognition, ontologies, data-mining, text-mining, automatic authorship identification, pattern analysis, optical character recognition, advanced statistics etc.) for diplomatic research?
- Have digital charter copora developed new research interests?
- Are there old research questions to be tackled by the digital technologies and digital charter corpora?
- Which well establish methods can't be accelerated by digital technologies?
- How far the internet the has changed scholarly communication in diplomatics?
- How you shape digitization projects of charters to meet research needs?

Abstracts

Nicolas Perreaux

(UMR 5594 ARTeHIS – Université de Bourgogne)

«From accumulation to exploration: experiences and proposals for indexing and for the use of diplomatics databases»

Keywords: Data and Text-Mining; automatic indexation; categorization; clustering; epistemology; factorial analysis; documentary production.

The mass of data accumulated through the growing digitization of old editions is just waiting to be explored. In fact, for several decades now, medievalists have at their disposal remarkable digitized databases, whose contents are now about to revolutionize our knowledge of the Middle Ages and this, at more or less short term. In this race to the accumulation of data, diplomatists are not outdone since several of the most impressive historical-text-databases were created by and for the actual leaders of the discipline (base des originaux de l'Artem, the *Chartae Burgundiae Medii Aevi*, the *Codice diplomatico della Lombardia medievale*, the Deeds Project, the *Monasterium* site, etc.).

However, as Benoît-Michel Tock recently pointed out[1], it is clear that the exploitation of the vast *corpora* – or rather should we say of this vast *corpus* -, remains largely to be done, while trials in this field are left in an embryonic state, and this, despite the goodwill and interest (but, we must admit it, also mainly occasional) of medievalists for those new technologies. Thereby, often put off by the difficult management and indexation of such a mass of documents, researchers generally return to more traditional methods, while using these *corpora* in parallel, but only as simple quarry of data. This lack of global method leads me to think, after the famous epistemologist Thomas S. Kuhn[2], that diplomatics is facing a kind of scientific crisis. Initially, this paper will be the opportunity to examine, following Alain Guerreau[3], the origin of the structural obstacles that still prevent the use of these now widespread resources. Nevertheless, we will defend the

[1] «Mais curieusement, celles qui existent [...] ne sont pas encore très utilisées. Disons le franchement : elles sont sous-utilisées, et n'ont pas, ou pas encore révolutionné la diplomatique comme elles auraient dû le faire. » dans B.-M. TOCK, «L'apport des bases de données de chartes pour la recherche des mots et des formules », in G. VOGELER (dir.), *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*, Köln-Weimar-Wien, 2009, p. 283-293, ici p. 283.

[2] KUHN Thomas Samuel, *The Structure of Scientific Revolutions*, The University of Chicago, Chicago, 1962.

[3] GUERREAU Alain, *L'avenir d'un passé incertain*, Seuil, Paris, 2001.

idea that – for the practical side of the problem at least - solutions exist, but are based on personal handling, by historians, of statistical tools and computer programming.

We thus propose to present a thesis in progress about the exploitation of these multiple *corpora*[4], and first of all the process that brought together nearly 150 000 charters in text mode in a single set, under the software Philologic (developed by Mark Olsen and his team at the University of Chicago: http://sites.google.com/site/philologic3/). Especially, we will introduce a system for the automatic indexation of charters by « author », based on Data and Text-Mining, but also an approach to more effectively manage the problem of time ranges in these databases. In fact, we all know that the correct and massive indexation of these documents is a prerequisite for anyone wishing to exploit these databases in a satisfactory manner. Indeed, one of the recurring obstacle against the use of this mass of texts remains the difficulty of distinguishing, for example, diploma from «simple» charters. The automatic classifications methods (classification / clustering), based on artificial intelligence, can now perform some of those sorts automatically. In this case, we will present a script, based on several algorithms (including the Support Vector Machine, the Naives Bayes and another one developed for this situation), that can now tag over 90% of papal documents and diplomas. Written in Perl, using several already existing modules, we will explain how this script can not only label charters but also inject the results of these experiments into Philologic, so that these categories can be used by diplomatists when it comes to compare the vocabulary of several types of documents.

Of course, one can imagine that this method could be extended to the issue of undated charters: documents would then be classified using the same tools. The main advantage of these techniques is that they only need – once the decision tree is written and data-mining algorithms implemented-, the creation of a database of training files (which is relatively simple to do by now since we already got a quite remarkable mass of indexed charters, as we already pointed it out). However, this technological phase obviously cannot stand on its own. In order to remain as concrete as possible, we will present a set of experiences already made about the contents and the dynamics of documentary production at the scale of medieval Europe from the IXth to the mid XIth century. This experiment, using factorial analysis and lexical statistics, aims, with the help of our database, to highlight certain features concerning the writing and the dynamics of charters. The purpose of this global approach is of course to bring to light, or even to model,

[4] «L'écriture du monde. Perception, catégorisation et appropriation de l'environnement dans les sources numérisées du Moyen Âge (VIIIe-XIIe s.) : une approche informatique, sémantique et statistique ». Thesis under the direction of Eliana Magnani and Daniel Russo, with the help of Alain Guerreau, in progress at the UMR CNRS 5594 ARTeHIS, University of Burgundy.

structures – as stated in the conference call for papers – remained invisible to the naked eye. Then I will defend the idea that charters production, even on a very large scale, has a structure and a strong social sense, above all articulated in large and clearly distinct groups of regions.

«De l'accumulation à l'exploitation: propositions et expériences pour l'indexation et l'utilisation des bases diplomatiques numérisées»

Mots-clés : Data et text-mining ; indexation automatique ; catégorisation ; clustering ; épistémologie ; analyses factorielles ; datation des actes ; production documentaire.

La masse de données accumulées à travers la croissante numérisation des éditions anciennes ne demande qu'à être exploitée. Depuis maintenant plusieurs décennies en effet, les médiévistes disposent de bases de données numérisées remarquables, dont le contenu est désormais propre à révolutionner nos connaissances concernant le Moyen Âge et ceci à plus ou moins cours terme. Dans cette course à l'accumulation de données, les diplomatistes ne sont pas en reste puisque plusieurs des bases parmi les plus remarquables ont été constituées par et pour les tenants de la discipline (base des originaux de l'Artem, les *Chartae Burgundiae Medii Aevi*, le *Codice diplomatico della Lombardia medievale*, le Deeds Project, le site *Monasterium*, *etc.*).
Pour autant, ainsi que le faisait encore récemment remarquer Benoît-Michel Tock[5], force est de constater que l'exploitation de ces vastes *corpus* – ou plutôt devrions-nous dire de ce vaste *corpus* – reste encore largement à faire, les entreprises dans le domaine restant pour le moment embryonnaires et ceci malgré la bonne volonté et l'intérêt notable (mais aussi, il faut bien l'admettre, ponctuel) des médiévistes pour ces nouvelles technologies. Ainsi, souvent rebutés par la difficile gestion d'une telle masse documentaire, les chercheurs retournent souvent à des méthodes plus traditionnelles, employant, certes, en parallèle ces *corpus*, mais comme de simples carrières de données. Cette absence de méthode globale nous conduit à penser, après le célèbre épistémologiste Thomas S. Kuhn[6], que la diplomatique doit faire face à ce qu'il nomme une « crise scientifique ». Dans un premier temps, cette communication sera donc l'occasion de s'interroger, à la suite d'Alain Guerreau[7], sur l'origine des blocages structurels qui empêchent encore à l'heure actuelle l'utilisation massive de ces ressources néanmoins remarquables. On défendra pourtant l'idée que des solutions existent, mais qu'elles reposent avant tout sur une prise

[5] «Mais curieusement, celles qui existent [...] ne sont pas encore très utilisées. Disons le franchement: elles sont sous-utilisées, et n'ont pas, ou pas encore révolutionné la diplomatique comme elles auraient dû le faire.» dans B.-M. TOCK, «L'apport des bases de données de chartes pour la recherche des mots et des formules», in G. VOGELER (dir.), *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*, Köln-Weimar-Wien, 2009, p. 283-293, ici p. 283.

[6] T. S. KUHN, *The Structure of Scientific Revolutions*, The University of Chicago, Chicago, 1962.

[7] A. GUERREAU, *L'avenir d'un passé incertain*, Seuil, Paris, 2001.

en main personnelle, par les historiens, des outils statistiques et de la programmation informatique.

On se propose ainsi de présenter une thèse en cours autour de l'exploitation de ces multiples *corpus*[8], et tout d'abord la méthode qui a permis de regrouper près de 150 000 chartes en mode texte au sein d'un ensemble unique, sous le logiciel Philologic (développé par Mark Olsen et son équipe à l'Université de Chicago : http://sites.google.com/site/philologic3/). Surtout, on présentera un dispositif d'indexation automatique, basé sur le Data / Text mining, ainsi qu'une démarche visant à gérer plus efficacement le problème des fourchettes chronologiques dans ces bases de données. De fait, un des blocages récurrents face à cette masse reste la difficulté qu'il y a à traiter distinctement, par exemple, les privilèges et les simples chartes. Les méthodes de classifications automatiques (catégorisation / clustering), basées sur l'intelligence artificielle, permettent désormais d'effectuer certains de ces tris de manière automatique. Dans le cas présent, on présentera un *script* créé pour l'occasion, basé sur plusieurs algorithmes (comprenant le Support Vector Machine, le Naives Bayes, ainsi qu'un autre développé afin de répondre au problème précis des documents diplomatiques) qui permet d'étiqueter automatiquement plus de 90% des bulles et des diplômes. Écrit en Perl, recourant à des bibliothèques de fonctions déjà disponibles, on montrera que cette méthode ne permet pas uniquement l'étiquetage des documents, mais permet aussi de réintroduire les résultats de ces expériences dans Philologic, afin que ces catégories soient utilisables par les diplomatistes lors d'expériences visant à comparer le vocabulaire contenu dans telle ou telle typologie d'actes.

Bien entendu, on peut imaginer que cette méthode soit extensible à la question des chartes non datées, qu'il s'agirait alors de classer grâce à ces mêmes outils. L'avantage principal de cette méthode est qu'elle demande seulement – une fois l'arbre de décision imaginé et les différents algorithmes de data-mining implémentés -, la création d'une base de fichiers d'entraînements (une tâche relativement facile à accomplir puisque nous disposons dès maintenant d'une masse remarquable de chartes indexées, ainsi qu'on l'a déjà signalé). Néanmoins, cette phase «technologique» ne peut évidemment pas se suffire à elle-même. Afin de rester le plus concret possible, on présentera surtout une série d'expériences d'ores-et-déjà réalisées, concernant la dynamique de la production diplomatique à l'échelle de l'Europe médiévale, du IX^{e} au milieu du XI^{e} siècle. Cette expérience, recourant aux analyses factorielles et à la statistique lexicale, vise, à partir de

[8] «L'écriture du monde. Perception, catégorisation et appropriation de l'environnement dans les sources numérisées du Moyen Âge (VIIIe-XIIe s.) : une approche informatique, sémantique et statistique». Thèse sous la direction d'Eliana Magnani et de Daniel Russo, avec l'aide d'Alain Guerreau, réalisée au sein de l'UMR 5594 ARTeHIS, à l'Université de Bourgogne.

notre base de données donc, à mettre en lumière certaines caractéristiques relatives à l'écriture et à la dynamique des documents diplomatiques. Le but de cette démarche globale est bien entendu de faire apparaître, voire de modéliser, des structures – tout comme le précise l'appel à communication du colloque – restées invisibles à l'œil nu. On défendra alors l'idée que la production diplomatique, même à une échelle aussi vaste, possède une structuration et un sens social forts, avant tout articulés en grands ensembles régionaux nettement distincts.

Olivier Canteaut, Frédéric Glorieux
(École Nationale des Chartes)

An attempt at clustering some royal French letters (14th-15th centuries)

Nowadays, diplomatists have gained access to more and more digitized documents, which have become far too numerous for any human mind to try remembering or even reading them all. That is why statistics and computer tools have become essential for mastering this massive data. The French royal letters of late Middle Ages, because of their standardized canvas, constitute a tantalizing case study, as they can be analysed efficiently by linguistical statistics.

The corpus we are focusing on, namely the charters registered at the royal chancery from the beginning of the 14th century, is particularly wide. Unfortunately, only very partial editions of it are available ; though, the edition of the acts about Poitou by Paul Guérin (end of the 19th century) provides a consistent sample of the whole. Our analyses have focused on 300 documents of this edition (178 documents dated before 1333 and 137 between 1390 and 1403), which have been digitized (http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/) in order to be investigated. As the chancery uses both Latin and the vernacular, two sets of documents must be delineated according to the language used : 137 are in Latin, 178 in French.

Whatever the chronology and the language used, when one examines this corpus through the correspondance analysis and the agglomerative hierarchical clustering (AHC), one obtains rather homogeneous sets of documents, according to both diplomatic and legal criteria. The writs belonging to the corpus are easily discriminated from the charters ; and among the charters, the letters of remission, the confirmations and the safeguards are clustered in consistent series. Sometimes one can improve the results by preprocessing the texts, for example taking out the proper nouns or selecting certain parts of the diplomatic structure. Though such a preprocessing does not modify radically our results.

Though, interpreting this classification remains difficult as long as one does not have a global insight into the production of the chancery. That is why the formulary compiled by the notary

Odart de Morchesne at the beginning of the 15th century is so valuable. This formulary has been edited under a digitized form by Olivier Guyotjeannin and Serge Lusignan in 2006 (http://elec.enc.sorbonne.fr/morchesne). It provides 270 formulae which, as representative of the whole of the chancery production, can be fruitfully compared with the corpus of the Poitou documents. To facilitate this comparison, the editorial treatment, which differs from one philologist to the other, have to be standardized, for example by homogenizing the resolution of abbreviations, the use of accents, the words troncation, etc. Once these preprocessings have been done, correspondence analysis allows useful comparisons. They confirm that, from a lexical point of view, lots of acts such as remissions or confirmations are standardized as early as the beginning of the 14th century. Other categories of documents do change in the course of time. For example, there is a difference in the form the king gives to his « lettres de sauvegarde » : in the beginning of the 14th century, they were written as charters, whereas a century later, they appear under the form of letters patent in Odart de Morchesne's formulary. Other types of documents can hardly be classified because, although their vocabulary is rather homogeneous, their form is not constrained enough. This is especially the case of the documents which deal with the domain (gift, grant, selling, alienation in mortmain, etc.).

In conclusion, it sounds relevant to challenge the usual classification of the royal acts according to legal criteria by using automatic treatment. This shall help dating the standardization of their form, and also to analyze their changes. These statistical tools offer another advantage : they shall enable us to cluster enormous masses of digitized charters. As the results we obtained prove statistically robust, we dare say that detecting royal charters among a heterogeneous corpus seems possible ; these acts could even be classified automatically, at least those which are both common and standardized.

Essai de classification automatique des actes royaux français (XIVe-XVe siècles)

Le diplomatiste est aujourd'hui confronté à un nombre toujours croissant de documents édités sous forme électronique. S'orienter dans cette masse, l'appréhender alors qu'elle dépasse les capacités humaines de lecture et de mémorisation, constitue désormais un enjeu essentiel que les outils informatiques et statistiques peuvent contribuer à résoudre. A ce titre, les actes royaux français de la fin du Moyen Âge, par leur rédaction largement standardisée, s'avèrent être un terrain particulièrement propice à des expériences d'analyse grâce aux méthodes de la statistique lexicale.

Les expériences que nous menons reposent sur le très vaste corpus que constituent les chartes enregistrées à la chancellerie royale depuis le début du XIVe siècle. Celles-ci n'ont malheureusement fait l'objet que d'éditions très partielles; toutefois, l'édition des actes relatifs au Poitou, réalisée à la fin du XIXe siècle par Paul Guérin, offre un échantillon homogène de cet ensemble. Grâce à l'édition électronique d'une partie de ce travail (http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/), a été constitué un corpus de travail de quelque 300 actes du XIVe siècle (178 du premier tiers du siècle et 137 de la dernière décennie). Le bilinguisme latin-français de la chancellerie royale contraint toutefois à séparer ce corpus en deux sous-ensembles, formés respectivement de 137 actes en latin et 178 en français.

Dans les deux langues, et quelle que soit la période considérée, l'exploration du corpus par les méthodes de l'analyse factorielle des correspondances (AFC) et de la classification ascendante hiérarchique (CAH) permet de faire émerger des groupes d'actes relativement homogènes, tant sur le plan diplomatique que juridique. Les mandements présents dans le corpus se distinguent ainsi de l'ensemble des chartes ; parmi ces dernières, les lettres de rémission, les confirmations, les sauvegardes forment également des groupes cohérents. La classification obtenue peut parfois être affinée en procédant à des pré-traitements des textes (élimination des noms propres ou restriction à certaines parties du discours) ; ceux-ci ne modifient toutefois pas les classes observées précédemment.

Les typologies obtenues s'avèrent cependant peu explicites, en l'absence de toute vision synthétique de la production de la chancellerie. C'est cette vision que nous offre le formulaire de chancellerie réalisé au début du XVe siècle par le notaire Odart de Morchesne. Grâce à l'édition électronique de ce formulaire en 2006, par les soins d'Olivier Guyotjeannin et de Serge Lusignan (http://elec.enc.sorbonne.fr/morchesne), 270 formules représentatives de l'ensemble de la production de la chancellerie peuvent être confrontées avec le corpus des actes relatifs au Poitou. Certes, cette comparaison nécessite d'uniformiser, dans la mesure du possible, les pratiques philologiques des éditeurs de ces deux ensembles d'actes, que celles-ci touchent à l'accentuation, à la coupure des mots ou encore à la résolution des abréviations. Mais, moyennant de légers pré-traitements, un examen conjoint s'avère possible grâce à l'analyse factorielle des correspondances. Il confirme le caractère très stéréotypé, sur le plan lexical, de nombreuses catégories d'actes (rémissions, confirmations…) dès le début du XIVe siècle. D'autres types d'actes connaissent des inflexions dans leur rédaction au cours du siècle : ainsi le roi émet-il de nombreuses lettres de sauvegarde sous forme de chartes au début du XIVe siècle, mais, un siècle plus tard, elles n'apparaissent plus que sous forme de lettres sur double queue dans le formulaire d'Odart de Morchesne. D'autres catégories encore, en particulier celles touchant aux questions domaniales (dons, ventes, amortissements…), se révèlent rétives à toute classification automatique : toutes emploient un vocabulaire proche et suivent des canons rédactionnels bien moins figés et standardisés que le reste de la production de la chancellerie.

Il apparaît ainsi possible de réinterroger les typologies juridiques des actes royaux, afin de mieux saisir les processus de standardisation, mais aussi d'innovation, à l'œuvre dans leur rédaction. Dans le même temps, de telles expériences d'analyse statistique lexicale ouvrent la voie à des procédures de classification automatique de grandes masses d'actes numérisés. Au regard de la robustesse des résultats obtenus, il semble possible de mettre en œuvre un repérage automatique des chartes royales au sein d'un plus large corpus et une classification automatique de ces actes, du moins pour leurs types les plus courants et les plus standardisés. Toutefois, les documents diplomatiques constituent des textes courts qui poussent les procédures statistiques à leurs limites : la diplomatique pose de nouveaux problèmes à la statistique textuelle, menant à chercher d'autres algorithmes.

Michael Gervers, Gelila Tilahun

(University of Toronto)

"Statistical Methods for Dating Collections of Medieval Documents"

A primary objective of on-going research at the DEEDS Project (Documents of Early England Data Set) at the University of Toronto is to develop statistical methods for the dating of undated English private (as opposed to royal) charters, of which only 5% were issued with dates between 1066 and 1307. Researchers at DEEDS have developed a database of over 10,000 dated charters from the period and used it to recognize chronological differences in word order and vocabulary which, through the application of statistical methodology, have enabled us to establish a temporal ``footprint'' for undated charter sources. In this paper, we present two such statistical methodologies that rely on usage patterns of words and phrases, and on the notion of 'distances' between documents. Both methods are computer automated and use the DEEDS database as their source.

In the first method, we define 'distances' between dated and undated documents and use these distance measures in a kernel function as weights for the dates of the dated documents in a non-parametric regression setting. In the second method, we estimate, as a function of time, the probability of the occurrence of words and phrases which occur in undated documents. This estimation is based on dated documents through the use of non-parametric regression techniques for generalized linear models. By combining the estimated probability of the occurrence of words and phrases, we estimate the date of an undated document. These methods could also be adapted to a setting in which the features are ordered or unordered catagoricals (they could, for example, be used to identify a religious house that composed a charter or a geographical location from where a charter originates).

Els De Paermentier

(Ghent University)

Diplomata Belgica: analysing medieval charter texts (dictamen) through a quantitative approach

In a recently finished doctoral study on the charters and chancery of the count(esse)s of Flanders and Hainaut (1191-1244), a 'traditional' combined methodology of diplomatics, palaeography and prosopography was applied. However, the availability of the digital source collection *Diplomata Belgica* offered the opportunity to extend the existing diplomatic method of 'Stilvergleichung' elaborated by L. Delisle and Th. von Sickel, and refined by Walter Prevenier in the early 1960s, with a whole new dimension, namely that of a quantitative word approach ('word statistics') within a corpus of over more than 16,000 charter texts issued between 1191 en 1244 by all kinds of secular and ecclesiastical authorities from within the area of present-day Belgium and northern France. The existing traditional method of diplomatic analysis had been limited to a 'manual' comparison of the Latin protocol formula, and only juxtaposed the text of the charters issued by the count(esse)s with other texts from the archives of their recipients. The new research criteria and standards that were worked out, gathered into a so-called 'three step action plan of determination', made it possible for the first time also to draw the dispositive text parts into the analysis, and to examine them from a much more comparative and 'creative' perspective. Consequently, this 'modern' methodology was not only elaborated in order to find out the editorial origin of the comital charter texts. Gradually, it also offered new insights on the editorial traditions and 'innovations' within the chancery of the count(esse)s, on the extent to which this chancery tended to differentiate itself from other secular or ecclesiastical editorial areas during the period 1191-1244, and on the direct influence some important chancery clerks had on the organisation and the editorial customs within the administrative entourage of the count(esse)s. Furthermore, the results of the *dictamen* analysis turned out also to be very valuable for digital discourse analysis, in order to determine in what way the charter texts drawn up in the comital chancery were used by the count(esse)s as an instrument to strengthen their legal authority and power towards their subjects.

Robin Sutherland-Harris

(Centre for Medieval Studies, University of Toronto)

Applications of the DEEDS Database to Somerset Charters: Dating, Diplomatics, and Historical Context

The DEEDS (Documents of Early England Data Set) Project currently consists of over

10.000 digitised dated charters. While the aim of the database is to provide a means, via the application of statistical algorithms, for dating undated charters, there are other avenues of exploration available. The application of meta-data to the charters of the database encodes information on people and places appearing in the documents.

A detailed demarcation of the diplomatic parts in each charter is also underway. The combined full resources of the DEEDS Project make available a wealth of information on private charters of medieval England.

My dissertation looks at how diplomatic materials reflect administrative trends in late twelfthand early thirteenth-century Somerset. This paper is constructed in part as a demonstration or test-run of the applicability of DEEDS Project resources to archival material external to the database, and in part as a preliminary investigation into a particular group of documents relevant to my research. Currently deposited at the Somerset Record Office are a number of loose charters, many with original seals. These documents involve transfers of land or rights over land at various levels of lay and ecclesiastical administration, and are for the most part only very loosely dated. Here a close study of those charters purported to be from from the reigns of Richard I and John (1189-1216) takes advantage of the resources made available by the DEEDS Project to analyse local administrative instruments, which can later be compared with those of

other levels of government. By applying DEEDS algorithms for the dating of undated charters we can not only move towards pinpointing a date for each individual document, but we can also observe exactly how linguistic features over a narrow time-span in these Somerset charters compare with evidence from documents drawn from across England and from a much broader time-span. More general trends visible in the material contained in the DEEDS database can be set against the un-digitised charters so as to illuminate particular points relating to diplomatic phrasing, unusual word use, and orthography. By making use of data on individuals and locations as they appear in both digitised and un-digitised documents, we can make expected (or unexpected) connections. The practical applications of the DEEDS database deepen our understanding of the temporal, diplomatic, and historical context of the Somerset charters. As these charters are not included in the DEEDS database, one of the fundamental question to be addressed here concerns the applications of digitised materials to the study of non-digitised documents. What can be said about the practical uses of digital medieval charters for those concerned with material external to digital databases?

Timo Korkiakangas

(University of Helsinki, Finland)

Challenges of the Linguistic Annotation of an Early Medieval Charter Corpus

In my paper I will view diplomatics from the point of view of a linguist. The paper discusses the challenges that the formulaic nature of medieval charters poses to a linguistic (syntactic) analysis of their language.

My Ph.D. research examines quantitatively the Latin noun declension system in a corpus of ca. 500 private charters from the 8th to 9th century Tuscany. The texts have been digitised from three copyright-free editions. I execute my analysis through a morphosyntactic annotation in the Perseus Latin Treebank online environment (Perseus Digital Library Project). Each word of the corpus will be provided with a lemmatic, morphologic and syntactic analysis in the form of an XML code. The queries will be run with the Annis search engine.

In a single charter many linguistically differing elements can be found. A crucial factor in any linguistic study of medieval charters is to distinguish the less formulaic or "freer" parts of a charter from the formulaic ones: these two parts present totally different linguistic realities. With the "freer" parts I mean the sections of charters that have been improvised by the scribe, while the formulaic sections rely on conventional formulae. I recognise that formulaicity is a continuum, but for ergonomic reasons I do not pay attention to the differing degrees of formulaicity in texts. I follow the rather dichotomic stand of Francesco Sabatini (1965) and Pär Larson (2000), instead.

A linguist cannot treat the language of the formulaic parts and the "freer" parts in the same way. In the formulaic parts, the general complexity and obscurity of the traditional chancery language has led the scribes to produce contaminational errors that are based on miscomprehensions or on a total lack of understanding of certain passages. The early medieval Italian scribes did not use formularies but composed the charters producing the required formulae from memory or copying appropriate passages from other charters (Schiaparelli 1933).

The "freer" parts are likely to represent spoken language more directly. A typical "freer" part is the disposition which contains the case-specific details of the legal act that has made the writing of the charter necessary. Nevertheless, even the disposition contains many formulaic expressions such as the dispositive verb itself, the possible *pertinentia* clause etc. Other places for "freer" sentences are, for example, *sanctio* clauses, inventories and coeval dorsal notes.

For my purposes, it will be enough to label the "freer" parts with a specific tag (and, probably, the subscriptions with another tag). I will not annotate isolated "freer" words, such as the ordinal numbers in the *datatio*, because they are usually not relevant for syntactic analysis. I have prepared myself for doing the annotation proper by hand, once I know which is "freer" and which is not. I am currently looking for ways to distinguish between the different parts with the help of computer. There seem to be several promising tools available that may be of help in detecting formulaic expressions.

The Janus n-gram intertextuality search engine, developed among others by Andrew Kane (Kane – Tompa 2010), finds overlaps between keywords and the *Manipulus florum* corpus of medieval Latin quotations. Janus can also be used for tracing the wordings of a certain formula inside my corpus. In consequence, the minimal and the maximal breadth of a formula can be defined and its basic form reconstructed. In the future, the detecting process can be automated.

Another possible tool is the n-merge method devised by the Multi-Version Documents (MVD) project promoted by Desmond Schmidt (Schmidt – Fiormonte 2010). The MVD technique is designed for finding out and clearly representing the divergences between texts that exist in multiple versions, e.g. between several manuscript witnesses. The application works as far as the texts exhibit at least 20% of similarity. In theory, the MVD technique allows all the charters to be compared with each other simultaneously. In its present state, the MVD system does not, however, cope with as many and as heavy parallel text files as those of my charter corpus.

A further aid for my study may be the dataset similarity measurement through diffusion maps, a relatively new method developed by my colleague Tuomo Sipola (Ph.D. student in Information Technology, University of Jyväskylä). At the moment, Sipola is adapting the method for finding out possible tendencies that could distinguish the formulaic passages from the "freer" ones. Given that with Annis structurally similar syntactic trees can be extracted from the treebank, it would also be useful to combine syntactic search to an n-gram approach. An allusion detection method of this type has been developed within the Perseus Project.

All the described tools only help in finding out formulae, not in annotating them in XML. A huge amount of interpretative work as well as of manual labour will be needed in spite of all the modern technical support. At their present stage of usability, the tools help me only in as much as they can teach me in uncertain cases which is formulaic and which is not.

References

Bamman, David – Crane, Gregory. 2008. 'The Logic and Discovery of Textual Allusion', *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data*, Marrakech (LaTeCH 2008).

Charter Encoding Initiative. http://www.cei.uni-muenchen.de/

The Janus Intertextuality Search Engine. http://web.wlu.ca/history/cnighman/page13.html

Kane, Andrew – Tompa, Frank. 2010. 'Janus: the intertextuality search engine for the electronic *Manipulus florum* project', *Literary and Linguistic Computing* 26, doi 10.1093/llc/fqr009.

Larson, Pär. 2000. 'Tra linguistica e fonti diplomatiche: quello che le carte dicono e non dicono', *La preistoria dell'italiano*. Atti della Tavola Rotonda di Linguistica Storica. A cura di József Herman e Anna Marinetti, 151–166.

Perseus Guidelines. Guidelines for the Syntactic Annotation of Latin Treebanks. http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf

The Perseus Ancient Greek and Latin Dependency Treebanks. http://nlp.perseus.tufts.edu/syntax/treebank/

Rio, Alice. 2009. *Legal Practice and the Written Word in the Early Middle Ages*: Frankish Formulae, c. 500-1000. Cambridge Studies in Medieval Life and Thought. Fourth Series 75.

Sabatini, Francesco. 1965. 'Esigenze di realismo e dislocazione morfologica in testi preromanzi', *Rivista di Cultura Classica e Medievale* 7, 972–998.

Sanga, Glauco – Baggio, Serenella. 1995. 'Sul volgare in età longobarda', *Italia settentrionale: crocevia di idiomi romanzi*. A cura di E. Banfi, G. Bonfadini, P. Cordin, M. Iliescu, 247–260.

Schiaparelli, Luigi. 1933. 'Note diplomatiche sulle carte longobarde, II: Tracce di antichi formulari nelle carte longobarde', *Archivio storico italiano* 19, 3–34.

Schmidt, Desmond – Fiormonte, Domenico. 2010. 'Documenti multiversione: una soluzione per gli artefatti testuali del patrimonio culturale / Multi-version documents: a digitisation solution for textual cultural heritage artefacts', *Intelligenza artificiale* 4, 56–61.

Sipola, Tuomo – Juvonen, Antti – Lehtonen, Joel. [forthcoming 2011] 'Anomaly detection from network logs using diffusion maps', *EANN/AIAI 2011*, Part I, IFIP AICT 363, 172–181. Ed. by L. Iliadis, C. Jayne.

Jeroen Deploige, Guy De Tré et al.

(Ghent University – BE)

When were medieval benefactors generous? Time modelling in the development of the database Diplomata Belgica

In the autumn of 2010, a team of both medieval historians and IT-specialists at Ghent University and at a number of partner institutions, amongst which most notably the Belgian Royal Historical Commission, started the project "Sources from the Medieval Low Countries (SMLC). A Multiple Database System for the Launch of Diplomata Belgica and for a Completely Updated Version of Narrative Sources" (Hercules Foundation Flanders, 2010-2015). SMLC will enable medievalists to gain free integrated online access to two different and completely up-to-date data collections: (1) Diplomata Belgica containing descriptions (and often full text and photographs) of some 33,000 charters composed in the Southern Low Countries before 1250, and (2) Narrative Sources, which offers a survey of all narrative sources written in the Low Countries before c. 1500.

Diplomata Belgica will become the most relevant component of SMLC for scholars of diplomatics. A primitive version of it is already known today as the CD-ROM Thesaurus Diplomaticus (Turnhout 1997). However, Diplomata Belgica will contain approximately a threefold number of descriptions of charters as well as a totally new relational database structure. This custom-designed software will allow us to handle our data in a way which coincides perfectly with our current research interests and needs and which offers the potential to develop more advanced methods of data mining and information retrieval in regard to future research questions. With this paper, we have three aims. We will first briefly present the project and its challenges. Then, in order to illustrate one of the many advantages of our new database technology, we will focus on procedures of time modelling of fuzzy information based on "possibility theory". The information historians have about the dates of the charters they are studying, is often marked by uncertainty, inaccuracy, vagueness, inconsistency or the lack of concrete data. By developing a semantic modelling approach, we will offer historians new means to undertake advanced searches for charters dating from specific periods. In the third part of our paper, we will explain how this technology might enable us to conceive new research questions about the promulgation of large numbers of medieval charters in relation to the different periods within the (liturgical) year. In brief, when were medieval benefactors especially generous?

Christian Emil Ore

(University of Oslo – NO)

Interlinking source text collections – a Norwegian example

Norway may differ from other countries in that few charters of Norwegian origin are preserved and the complete printed Diplomatarium Norvegicum, DN, (1846 – present) comprises less than 24 000 documents. Many of the transcripts are based on copy books in the Vatican and other sources. A newer series Regesta Norvegica, RN, with update source information and synopsis has been published and has reached 1419. DN has been retro-digitized 15 years ago. DN and RN are converted into XML and published on the web. They are currently being converted into TEI-P5. Medieval Norwegian Text Corpus (Menotec) is a project aiming at establishing a 1.5 million words corpus of transcriptions compliant with Menota (www.menota.org) including a 0.5 million words treebank. 20% of the corpus are based on charter texts. Since DN is not on the level of modern text edition, we use texts from a collection of high quality transcriptions (CT) done in 1970-1990. All charters in CT are described in the Regesta Norvegica series and opens for interlinking the texts. The current web-version of DN and RN is based on ad hoc markup. To give open access to the material, including the text corpus, it is necessary to use well defined formats like TEI and Menota. The metadata model both for the content and for the text versions is compliant to FRBR(oo) and CIDOC-CRM and thus easy to convert into RDF-format. This opens for interlinking with other sources collection for example via Linked Open Data.

Richard Higgins

(Durham University – UK)

Cataloguing medieval charters: a repository perspective

As custodians of a broad range of collections we require a system that enables cataloguing and presentation of related digital material that is flexible enough to cope with all materials. We have been using EAD as a data storage format since 1996, and in combination with our Fedora-Commons digital repository we have a powerful, adaptable tool. It is imperative that one system includes all our collections, so that enhancements and migrations apply to the whole and do not break or drop the more complex data. EAD has proven hugely adaptable and scalable, ranging from brief description of collections to the calendaring of individual charters. As one of hundreds of collections in our care, the archive of Durham Priory and Cathedral includes thousands of charters, as well as cartularies and a full range of medieval documents. EAD has been able to accommodate descriptions of all of these – even 3,000 seals. The digital repository also stores images and transcripts of the documents. It enables the association of description and image using index terms, hyperlinks, and RDF, producing a more permanent linking between data within the catalogue, and offering researchers the ability to make reliable citations of online representations of individual documents. This enables investigation of not just additional versions of the charter, but also other documents witnessed by the same parties and other common features. The proposed session would be a demonstration of the website, concentrating on how it presents individual charters in the context of the larger archive.

Pierluigi Feliciati

(Università degli studi di Macerata – IT)

Digital description and scanning of parchments and seals in the context of a national archival information system: the experience of SIAS

The international standards which support the provision of archival finding aids, first of all ISAD (G), did not face – correctly – the scientific description and analysis of single documents. They offer general rules to ensure the standardization of description logics and proper contextualization of each description unit both in the funds framework and inside the creation dynamics. The practical application of ISAD (G) to practical activities of description, except the Guidelines in 2001, was postponed to the working on formal models (especially for the EAD scheme) and to the development of individual descriptive software. In Italy, there was a parallel development – not always coordinated – of digitization projects and high level information systems. The SIAS project, developed in since 2003 by the General Directorate of Archives and then coordinated by ICAR, tried to bring together into one distributed platform the goal of building a national framework on the documentary heritage of Italian State Archives not forgetting the instances of descriptive activities at inventory level, by providing an appropriate software module with specific features including "special" documents like parchments and seals. In addition, the process of digital reproduction was supported by specific guidelines and adopting a model of administrative and structural metadata in order to ensure a stable link between digital images with appropriate descriptions. This presentation aims to highlight the SIAS scientific choices, with particular attention to the modules dedicated to parchments and seals, giving account after some years of the large qualitative and quantitative results and pointing to the current development strategies of the Central Archives.

Descrizione digitale e digitalizzazione di pergamene e sigilli nel contesto di un sistema informativo archivistico nazionale: l'esperienza del SIAS

Gli standard internazionali che supportano la predisposizione di strumenti di ricerca archivistica, e tra questi in particolare ISAD(G), non si sono addentrati – correttamente, peraltro – nelle questioni scientifiche relative alla descrizione analitica dei singoli documenti. Essi dettano piuttosto regole di carattere generale volte a garantire la standardizzazione delle logiche descrittive e la corretta contestualizzazione di ogni oggetto descrittivo sia nell'alveo dell'*universitas* documentaria che rispetto alle dinamiche di produzione. L'applicazione pratica di ISAD(G) alla descrizione "sul campo", a parte la felice parentesi delle *Guidelines* del 2001, è stata rinviata ai tavoli di lavoro sui modelli formali (soprattutto per lo schema EAD) e allo sviluppo dei singoli software descrittivi. In Italia, si è osservato lo sviluppo parallelo e non sempre coordinato da un lato di progetti di digitalizzazione e (nei casi migliori) relativa descrizione analitica, dall'altro di progetti di sistemi informativi generali, di livello "guida", anche molto raffinati. Il progetto SIAS, nato nella Direzione Generale degli Archivi quindi nell'ICAR dal 2003, ha cercato di riunire in un'unica piattaforma distribuita l'obiettivo di costruire un quadro nazionale aggiornato sul patrimonio degli Archivi di Stato con le istanze descrittive di livello inventariale, prevedendo un modulo software ad hoc dotato di specifiche funzionalità dedicate anche a tipologie documentarie "speciali" come le pergamene e i sigilli. Inoltre, il processo di riproduzione digitale è stato supportato da specifiche linee guida e dall'adozione di un modello di metadati amministrativi e strutturali così da garantire il legame tra le immagini digitali con adeguati supporti descrittivi. Questa presentazione è volta ad evidenziare le scelte a suo tempo effettuate, con particolare attenzione alle schede per pergamene e sigilli nel contesto del SIAS, dando conto degli ampi risultati qualitativi e quantitativi e accennando le attuali strategie di sviluppo dell'Istituto Centrale degli Archivi.

Francesca Capochiani, Chiara Leoni, Roberto Rosselli Del Turco

(Università degli studi di Pisa – IT/ Università degli studi di Torino – IT)

Open source tools for online publication of charters

Diplomatic texts consultation is an indispensable tool for historians and archivists. Their online availability offers maximum flexibility and dissemination, allowing the scholar to access this valuable material unimpeded by spatial or temporal barriers: some projects, such as The Electronic Sawyer (http://www.esawyer.org.uk/) and the encoding activities of the École Nationale des Chartes (http://www.enc.sorbonne.fr), show how it is possible offer high quality scientific texts on the Web based on an XML mark-up of the archival sources. Their creation, however, requires substantial resources: is it possible to digitize and to put online this type of document collections for personal research, or to benefit the whole Academic community, in a (relatively) simple and effective way? Furthermore, Web publishing is only effective if it allows easy document browsing and a way to perform powerful data mining of the resources it offers: which methods are best to allow easy text access and search? The first step, digitisation, can be done either converting documents in simple text format, or using a mark-up language such as XML, using the TEI encoding schemas or similar ones. Taking the XML mark-up path can slow down the conversion to a MRF (Machine Readable Format), but it is just a minor annoyance compared to the benefits granted by a semantic encoding.

Using an XML search engine such as eXist (http://exist-db.org/) or XTF (http://xtf.cdlib.org/) allows to take advantage of the semantic encoding accomplished in the previous step. Being able to perform complex queries is particularly important for a scholar studying charters and similar documents because an effective data mining of digital texts is central to her/his purpose. Once charters have been encoded and integrated in an XML database they can also be shown on-line by means of simple XSLT style-sheets, or using the relative functionality of the XML databases, or even through specific software. This paper aims to show how, through the use of open source software, single scholars or a small team of researchers can encode a corpus of documents using the TEI standard, publish it on the Web, and provide a search engine so that a scholar may be able to effectively browse and study all the texts it contains.

Strumenti open source per la pubblicazione online di documenti diplomatici

La consultazione di testi diplomatici costituisce uno strumento di lavoro insostituibile per gli storici e gli archivisti. La loro disponibilità online offre il massimo della flessibilità e della diffusione, permettendo allo studioso di accedere a questo materiale prezioso senza barriere spaziali o temporali: alcuni progetti, come The Electronic Sawyer (http://www.esawyer.org.uk/) e l'attività della École Nationale des Charter (http://www.enc.sorbonne.fr), mostrano come sia possibile offrire testi di alta qualità scientifica sul web usando una codifica XML delle fonti. La loro creazione, tuttavia, richiede risorse non indifferenti: è possibile digitalizzare e mettere online questo materiale, per le proprie ricerche e per il beneficio della comunità accademica, in maniera (relativamente) semplice ed efficace? Inoltre una pubblicazione sul web è incompleta se non consente un'agevole consultazione e il data mining delle risorse offerte: come facilitare l'accesso e la ricerca all'interno dei testi? Per quanto riguarda la digitalizzazione, la scelta è fra il formato testo semplice e l'uso di un linguaggio di mark-up come XML, seguendo gli schemi di codifica TEI (http://www.teic.org/) o altri simili. L'uso di XML può rallentare la conversione in formato elettronico, ma si tratta di un problema secondario rispetto ai numerosi vantaggi garantiti dalla codifica semantica. L'uso di un motore di ricerca XML come eXist (http://exist-db.org/) o XTF (http://xtf.cdlib.org/) permette di mettere a frutto la marcatura semantica portata a termine nella fase precedente. La possibilità di effettuare ricerche complesse è particolarmente importante in quanto lo studioso di documenti diplomatici è interessato a un efficace datamining dei testi digitali. I documenti così digitalizzati e integrati in un database, infine, possono essere visualizzate sul web per mezzo di semplici fogli di stile XSLT, o usando le funzionalità offerte dal database, o ancora per mezzo di software dedicato. Questo intervento si propone di mostrare come, grazie all'uso di software open source, il singolo studioso o un piccolo team di ricercatori possa digitalizzare un corpus di documenti usando il formato TEI, pubblicarlo sul web e inserire nell'interfaccia un motore di ricerca in modo da offrire allo studioso tutto il necessario per studiare e consultare i documenti desiderati.

Antonella Ghignoli
(Università degli studi di Firenze – IT)

The "Italia Regia" Project

The scientific coordinators of this project are François Bougard (Université Paris X- Nanterre), Wolfgang Huschner (Universität Leipzig) and Antonella Ghignoli (Università di Firenze). It is an online system through which digitalizing and putting online documents promulgated by the royal power during the *Regnum Italiae* in the High Middle Age (VII-VIII century). If you consider the objects we want to describe, the system concept has been implemented through a *Database management system* presenting a linked data structure. The system allows to describe separately objects and their possible relations. The information system allows to manage different entities: prosopography cards divided into different type of people and bodies; documents cards divided into placita and diploms; bibliography cards. You will find attached a system general pattern with the main relations. The project is made by a management environment (http://lartte.sns.it/scriba) which allows to enter and modify information and by an editing environment (http://www.italiaregia.it) allowing to put online all the validated information entities. Users can interact with the system through the browser they prefer (Internet Explorer, Netscape, etc.) via web or a LAN. To be more precise, through an HTML pages interface users communicate from their browser to the system through an http server (Apache) asking it to do a series of cgi-bins (in Perl language). The cgi-bins communicate with the DB (MySQL) reading and writing data in relation with cards (prosopography cards, placita, diploms, users) when an authorized user ask for them. The system is now operating on Lynux Suse platform. On the whole 127 cards of diploms have been entered till now in the Data Management System (distributed by diocese in such a way: Arezzo: 35, Chiusi 13, Fiesole: 3; Florence: 20; Lucca:34; Pisa: 8; Pistoia: 2; Siena: 5; Volterra: 7) and 57 cards of placita (distributed by diocese in such a way: Arezzo: 1; Lucca: 55; Pisa: 1); together with an at the moment uncertain number of prosopography and bibliography cards generated within the relation structure. At the moment there are not many questionable cards and relations by external (from the Italia Regia project researchers group) users in the editing environment, because most of them have not been validated yet. This operation allows them to be edited. A lot of research is still ongoing in fact, even if they are in progress. Some have been accomplished and presented to the first meeting of the *Italia Regia* research group in Leipzig (2009, december 11-12) over the theme *Europäische Herrscher und die Toskana im Spiegel der urkundlichen Überlieferung* (which records are to be printed).

Il progetto "Italia Regia"

I coordinatori scientifici del progetto sono François Bougard (Université Paris X- Nanterre), Wolfgang Huschner (Universität Leipzig) a Antonella Ghignoli (Università di Firenze). È un sistema on line per la digitalizzazione e la restituzione via web dei documenti emessi dal potere regio nel *Regnum Italiae* nell'alto e pieno medioevo (secoli VII-XII). Il modello concettuale del sistema, tenuto conto degli oggetti da descrivere, è stato implementato mediante un sistema di gestione dei dati (*Database management system*) a struttura relazionale, che permette la descrizione separata degli oggetti e le loro possibili relazioni. Il sistema informatico permette di gestire varie entità: schede prosopografiche, distinte per tipologia in persone ed enti; schede documenti, distinte in placiti e diplomi; schede bibliografiche. Si allega uno schema generale del sistema con le principali relazioni. Il progetto si compone di un ambiente di gestione (http://lartte.sns.it/scriba) che permette di inserire e modificare le informazioni e di un ambiente di pubblicazione (http://www.italiaregia.it) che permette la pubblicazione on line di tutte le entità informative validate. Gli utenti possono interagire con il sistema utilizzando un qualsiasi browser (Internet Explorer, Netscape, ecc.) tramite Internet o attraverso una rete locale. Più precisamente: l'utente attraverso una interfaccia costituita da pagine HTML, dal proprio browser comunica con il sistema attraverso un server http (Apache) chiedendogli di eseguire una serie di cgi-bin (scritti in linguaggio Perl). I cgi-bin colloquiano con il DB (MySQL) leggendo e scrivendo i dati relativi alle schede (prosopografiche, placiti, diplomi, utenti) quando un utente che ne abbia l'autorizzazione ne fa richiesta. Attualmente il sistema è operativo su piattaforma Linux Suse. Per la provincia Etruria sono state inserite finora nell'ambiente di gestione complessivamente 127 schede di diplomi (così distribuite per diocesi: Arezzo: 35; Chiusi: 13; Fiesole: 3; Firenze: 20; Lucca: 34; Pisa: 8; Pistoia: 2; Siena 5; Volterra: 7) e 57 schede di placiti (così distribuite per diocesi: Arezzo: 1; Lucca: 55; Pisa: 1) insieme a un numero imprecisabile, al momento attuale, di schede prosopografiche e bibliografiche generate all'interno della struttura di relazione. Sono ancora poche, al momento attuale, le schede e le relazioni interrogabili nell'ambiente di pubblicazione da utenti esterni al gruppo dei ricercatori del progetto Italia Regia, perché la maggioranza di esse attende ancora di essere validata, operazione che ne determina automaticamente la pubblicazione sul web: molte ricerche, infatti, sono ancora in corso, benché in uno stato di forte avanzamento; diverse invece sono giunte compimento e sono state presentate al primo incontro congressuale del gruppo di ricerca di *Italia Regia* tenutosi a Leipzig (11-12 dicembre 2009) sul tema *Europäische Herrscher und die Toskana im Spiegel der urkundlichen Überlieferung*, i cui atti sono in corso di stampa.

Camille Desenclos, Vincent Jolivet
(École nationale des chartes, Paris – FR)

Diple, proposals for the convergence of XML/TEI schemas adapted to the edition of diplomatic sources

Since 2002, the Ecole des chartes has put time and effort into a program of electronic editions of historic sources, principally diplomatic ones, and has extended a long tradition of critical edition. As early as July 2003, the choice of TEI seemed obvious, so that we didn't have to multiply the namespaces while encoding editions, including editions of non diplomatic sources. If the choice of TEI proved excellent – TEI success can testify on behalf –, it wasn't well mastered yet, probably because of an incomplete understanding of its nature: not really a DTD but a base which can be used for defining various DTD adapted to specific projects. This confusion, added to the unavoidable trials and errors of the first experiences, led in our corpora to the multiplication of encoding solutions for the same needs. While the choice of TEI was motivated by an interoperability requirement, it has led to the multiplication of not very interoperable schemas in the same namespace. This paradox can still be found within many projects, restricts the possibilities offered by the digitalisation of huge corpora and induces to doubt of the pertinence of the TEI initiative: at the beginning of his project, the novice has to confront himself to the encoding guidelines and can be discouraged by the 500 documented elements. The publishing (HTML export, ePub, LaTex) is getting tiresome and expensive when it's about to write new XSL for each new corpus. Actually online editions aren't always impeccable and don't allow institutional acknowledgements which are as delicate as essential. As for publishing, writing generic exportprograms for statistic or lexicometric softwares is getting impossible and restrains the possibilities of scientific exploitations of corpora. We are thus losing almost the whole benefit of using a shared standard and are out of the scientific prospects offered by the crossed exploitation of huge corpora. By starting up again from this heterogeneous state, we have been able to look at this from all sides and try to resolve these difficulties. The method is well known, its principle simple but its implementing requires patience: writing schemas. The effort to formalize our editorial practices has been implemented within the development of Diple, a software for publishing and processing XML/TEI files. This paper's aim is to present those schemas dedicated to the critical edition of diplomatic sources and our documentation methods. There is a proposal to improve the convergence of the editorial practices for diplomatic corpora, facilitate a real interoperability and open new horizons for massive processing in these corpora. The systematic encoding of a critical edition involves specifying precisely the typology of its components. This work of definition cannot always be found within paper publications as their ancient norms are considered as obvious. However, as we are aware of the fact that we are adapting a French diplomatic and philologic tradition, we didn't start from scratch but are trying to translate the *Conseils pour l'édition des textes médiévaux* (dir. O. Guyotjeannin et F. Vieillard, Paris: CTHS ed., 2001) into Relax-NG schemas from concrete cases found in our editions (http://elec.enc.sorbonne.fr).

These schemas (http://developpements.enc.sorbonne.fr/diple/schema/) are conceived as a library of named patterns dedicated to the encoding of the components of the critical editions (tradition, critical apparatus, dates, etc.) as well as some useful proposals for encoding some diplomatic patterns (e.g. diplomatic formulas). By associating systematically to its schemas some functionalities to exploit them (HTML display, various exports), Diple facilitates the corpora uploading, resolves the display problems of recurrent text structures and allows the editor to concentrate on to this corpus specificities. Thus the editor writes his own schema by gathering the needed patterns and by adding the other patterns useful to his scientific project (e.g. diplomatic, palaeographic or linguistic markup). Using the XREM module in Diple allows the editor to view these schemas as a textual and browsable documentation. Such documentation is a useful flag for the scientific editor who isn't necessarily comfortable with XML. It contains the encoding rules, clarifies the scientific project and above all allows, thanks to the XML validation, to be sure of the conformability to these rules. From now on, we no longer have a schema on one side, the guidelines on another but a clear single file: the Relax-NG schema, which efficiently declares rules and compels us to respect them through the validation process. In our opinion such a practice, in addition to be a necessary condition to files exchange, provides materials for a reflexion about the institutional acknowledgment of the scholarly electronic editions and ensures quality of edition.

Diple, propositions pour la convergence de schémas XML/TEI dédiés l'édition de sources diplomatiques

L'École des chartes est investie depuis 2002 dans un programme d'édition électronique de sources historiques, principalement diplomatiques, prolongeant ainsi une longue tradition de l'édition critique. Dès juillet 2003, le choix de la TEI s'est imposé afin de ne pas multiplier les espaces de noms pour l'encodage d'éditions qui ne se limitent pas aux seules sources diplomatiques. Si ce choix s'est avéré excellent – le succès de la TEI en témoigne aujourd'hui –, il n'était pas encore bien maîtrisé, sans doute à cause d'une compréhension imparfaite de ce qu'est véritablement la TEI : non pas à proprement parler une DTD, mais plutôt un socle utile à la définition de différentes DTD, dédiées à des projets spécifiques. Cette confusion, conjuguée aux inévitables tâtonnements des premières expérimentations, s'est traduite dans nos corpus par la multiplication des solutions d'encodage arrêtées pour un même besoin. Alors même que le choix de TEI est guidé par un impératif d'interopérabilité, il a pu conduire à la multiplication de schémas peu interopérables, dans un même espace de noms. Cette situation paradoxale prévaut dans de nombreux projets aujourd'hui encore, bride les possibilités offertes par la numérisation de vastes corpus et conduit à douter de la pertinence même de l'initiative TEI : les quelques 500 éléments documentés des *Guidelines* peuvent décourager le néophyte qui doit s'y confronter au commencement de son projet. La publication (export HTML, ePub, LaTeX) s'avère fastidieuse et coûteuse quand il faut écrire de nouvelles XSL pour chaque nouveau corpus. Par conséquent, les éditions en ligne ne sont pas toujours irréprochables, ce qui ne conforte pas une reconnaissance institutionnelle aussi fragile qu'indispensable. Comme pour la publication, l'écriture de programmes génériques d'export vers des logiciels de traitements statistiques ou lexicométriques est rendue impossible, bridant les possibilités d'exploitation scientifique des corpus constitués. Nous perdons donc presque tout le bénéfice de l'usage d'un standard partagé, nous privant des perspectives scientifiques offertes par l'exploitation croisée de très vastes corpus. La reprise d'un existant hétérogène à l'École des chartes a été l'occasion d'une remise à plat avec pour objectif de lever l'ensemble de ces difficultés. La méthode est bien connue, son principe simple, mais sa mise en oeuvre requiert de la patience : la rédaction de schémas. Cet effort de formalisation de nos pratiques éditoriales a été mené dans le cadre du développement de Diple, un logiciel de publication et de traitement des fichiers XML/TEI. Cette communication a pour objet de présenter ces schémas dédiés à l'édition critique de sources diplomatiques et nos méthodes de documentation. Ces schémas sont une proposition pour améliorer la convergence des pratiques d'éditions des corpus diplomatiques, faciliter une réelle interopérabilité et ouvrir la voie au traitement de masse de ces corpus. L'encodage systématique d'une édition critique contraint à préciser finement la typologie de ses composantes. Ce travail de définition n'est pas toujours fait dans les publications papiers dont les norms déjà anciennes sont considérées comme des évidences. Pour autant, conscients d'adapter une tradition diplomatique et philologique française, nous ne sommes pas partis de rien, nous efforçant de traduire les *Conseils pour l'édition des textes médiévaux* (dir. O. Guyotjeannin et F. Vielliard, Paris : CTHS éd., 2001) en schémas Relax-NG, en nous appuyant toujours sur les cas concrets présents dans nos éditions (http://elec.enc.sorbonne.fr). Ces schémas (http://developpements.enc.sorbonne.fr/diple/schema/) sont conçus comme un réservoir (*library*) de motifs nommés (*named patterns*) dédiés à l'encodage des composantes de l'édition critique diplomatique (tableau de la tradition, apparat critique, datation, etc.). Y figurent également quelques propositions utiles pour l'encodage de certains motifs diplomatiques (repérage des parties du discours). Diple, en associant systématiquement à ses schémas des fonctionnalités pour les exploiter (affichage HTML, exports divers), facilite la mise en ligne des corpus, résout les problèmes de présentation des structures textuelles récurrentes et permet à l'éditeur de se concentrer sur les particularités de son corpus. L'éditeur compose ainsi son propre schéma en assemblant les motifs (*pattern*) utiles à l'édition et en y ajoutant ceux utiles à son projet scientifique (e.g. encodage diplomatique, paléographique ou linguistique).

Daniel Piñol Alabart

(Universidad de Barcelona – ES)

Proyecto ARQUIBANC. Digitalización de archivos privados catalanes: una herramienta para la investigación.

En Cataluña se conserva un notable patrimonio documental privado que es fundamental para estudiar la historia del país desde diferentes puntos de vista. Estos archivos privados lo configuran importantes archivos patrimoniales, archivos de empresas, archivos de familias o de personajes importantes. La mayoría de ellos se guarda en manos de sus propietarios, aunque hay algunos que están en instituciones archivísticas públicas, donde han ingresado por donación, compra o depósito. Uno de los problemas que revisten estos fondos documentales es, generalmente, el difícil acceso para los investigadores, aunque los que se encuentran en archivos públicos existen los instrumentos adecuados para poder acceder a ellos. Atendiendo a esta situación el proyecto ARQUIBANC, gestionado desde el Departamento de Historia Medieval, Paleografía y Diplomática, de la Universidad de Barcelona, tiene como objetivo digitalizar alguno de estos archivos para facilitar el acceso a investigadores, historiadores, diplomatistas, paleógrafos o archiveros. Con la presentación del proyecto se incluye un análisis de los problemas que conlleva el trabajo con documentación privada y los problemas que se generan en torno al proceso de digitalización. Sobre todo se atiende a la cuestión de si es necesario digitalizar documentos, qué documentos se digitalizan y cuáles no, y si la digitalización es un medio o un fin. Los documentos digitalizados se insertan en una base de datos con la descripción de cada uno de ellos y la imagen correspondiente. Para poder llevar a cabo la descripción archivística es necesario un estudio individualizado de cada documento, incluyendo estudios diplomáticos y paleográficos, así se puede dotar de información cada una de las fichas introducidas en la base de datos. Esta, además, se aloja en una web que incluye resultados del proyecto de investigación, herramientas útiles para los investigadores y listados de bibliografía sobre el tema del proyecto. Se presta atención también a los problemas relacionados con el trabajo con bases de datos, con su manejabilidad, con el acceso de los usuarios y la búsqueda de los documentos y la información de cada uno de ellos. Son problemas que se intentan solucionar de forma paralela al proceso de estudio y trabajo con los archivos y documentos privados que llegan a nuestras manos. Todo ello con el objetivo último de poner al alcance de la comunidad científica algunos archivos privados que, de otra manera, quedarían inaccesibles a los investigadores, a los profesores y profesoras y a los alumnos que ya pueden realizar trabajos de investigación con esta documentación.

Paul Bertrand, Maria Gurrado

(IRHT - CNRS)

Medieval Paleography And Formal Quantitative Analysis A Research About A Group Of Medieval «Quittances» (County Of Flanders, 1275-1325)

The quantitative approach to medieval documents does not know anymore the same interest as years ago since when Ezio Ornato, Carla Bozzolo, Peter Rück, Frank Bischoff and others considerably cleared a true virgin forest. The great methodological difficulties they faced have been widely described particularly in the *Gazette du Livre Médiéval*, with an awful effect widened by the declining interest in the statistics applied to History. There are a little works about the quantitative approach about medieval manuscripts, a fortiori, in writing.

The emergence of a certain number of research projects in digital humanities allows to hope for a renewal of these approaches about the quantitative codicology. On a certain regard this research takes part in it. It analyzes the formal structures present in some medieval manuscripts through the semi-automatic measures with the help of a plug-in software inserted over an application-view of numerical images. This open source and open access software is now a prototype version called *Graphoskop*. A creation by Maria Gurrado who used it to analyze the features of the cursive script within a group of French manuscripts between the 13th and 15th centuries, in her archival paleography thesis she recently discussed at the *École Nationale des Chartes*, headed by M. Smith.

The paleographic phenomenon about the cursive script development in the 13th-15th centuries is a very serious and important scientific matter regarding the history of legal texts, concerning particularly and in the widest sense the Diplomatics documents (until it is called a little rapidly «ordinary writing»). After studying the sample of the dated manuscripts, the resonance of which will be soon done by a publication by Maria Gurrado, both of us have considered the importance of applying the quantitative analysis techniques about the writing measures to a Diplomatics writing sample. A lot of aspects have already been considering to an ongoing global research, particularly a part concerning the writing of registers and bookkeeping.

The part we want now to treat concerns the charters writing, specifically the «lower ranking» charters, such as receipts and bills collected by the count of Flanders in his accounting archives all dating from 1275 to 1325, is a key moment of the birth of the cursive documentary. All the receipts make a coherent whole collected and saved without any other visible destructions since the end of the 13th century. Over 500 receipts about 200 have been chosen (for reasons of numerical and picture quality) and their writing have been the subject of several automatic measures.

These measures have been entered in a spreadsheet and correlated to other codicological measures taken from the original rather than the Diplomatics and documents data. The paper aims to present the results of our quantitative paleography research about the cursive script in these minor charters. We will try to understand how the Diplomatics cursive script emerged and developed in these documents to better understand this particular kind of quittance which is not again well known in medieval Diplomatics. The expose will be accompanied by methodological reflections about this research combining together digital humanities, paleography and Diplomatics.

Paléographie médiévale et analyse formelle quantitative. Enquête sur un ensemble de « quittances » médiévales (comté de Flandre, 1275-1325)

L'approche quantitative de l'écrit médiéval ne connaît plus l'engouement qu'elle suscitait il y a quelques années, lorsque Ezio Ornato, Carla Bozzolo, Peter Rück, Frank Bischoff et bien d'autres ont défriché remarquablement une véritable forêt vierge. Les grandes difficultés méthodologiques qu'ils ont rencontrées ont été largement décrites, notamment dans la Gazette du Livre Médiéval, avec un terrible effet, amplifié par le déclin de l'intérêt pour les statistiques appliquées à l'Histoire : bien peu de travaux sont désormais entrepris autour de l'approche quantitative du manuscrit médiéval, a fortiori de l'écriture.

L'émergence de nombre de projets de recherche en *digital humanities* permet d'espérer un renouveau de ces approches de codicologie quantitative. L'enquête présentée ici y participe, d'une certaine façon. Elle vise à analyser les structures formelles des écritures rencontrées dans des manuscrits médiévaux, par le biais de mesures prises de manière semi-automatique à l'aide d'un logiciel inséré en « plugin » sur une application de visualisation d'images numériques. Ce logiciel, open source et libre d'accès, actuellement en version prototype, est appelé Graphoskop. Il a été conçu par Maria Gurrado qui l'a utilisé pour analyser les traits de l'écriture cursive dans un ensemble conséquent de manuscrits datés sur l'espace français, entre le XIIIe et le XV^{e} s, dans sa thèse d'archiviste paléographe tout récemment soutenue à l'Ecole nationale des Chartes, sous la direction de M. Smith.

Le phénomène paléographique du développement de l'écriture cursive au cours des XIII^{e}-XV^{e} est un enjeu scientifique majeur pour l'histoire des pratiques de l'écrit, et notamment de l'écrit diplomatique au sens le plus large (jusqu'à ce que l'on appelle un peu rapidement «l'écriture ordinaire»). Après l'étude de l'échantillon des manuscrits datés, dont une publication se fera rapidement l'écho par Maria Gurrado, nous avons tous deux jugé important d'appliquer les techniques d'analyse quantitative des mesures d'écritures à un échantillon d'écritures diploma-

tiques. Plusieurs volets sont d'ores et déjà prévus pour une enquête globale en cours, dont notamment un volet dédié aux écritures de registres et de comptabilités.
Le volet dont nous vous entretenons ici touche aux écritures de chartes, plus exactement de chartes «de rang inférieur», des quittances et reçus accumulés par le comte de Flandre dans ses archives comptables, datant toutes d'entre 1275 et 1325, soit un moment clé pour la naissance de la cursive documentaire. Toutes ces quittances forment un tout cohérent, rassemblé et conservé sans destructions ultérieures apparentes depuis la fin du XIIIe s. Sur les 500 quittances, 200 environ ont été sélectionnées (pour des raisons de qualité numérique des images) et les écritures qu'elles portaient ont fait l'objet de mesures automatiques nombreuses. Ces mesures ont été introduites dans un tableur et corrélées avec d'autres mesures codicologiques prises sur l'original ainsi qu'avec des données diplomatiques et documentaires. L'exposé vise ici à présenter les résultats de l'enquête de paléographie quantitative menée autour de la cursive dans ces chartes mineures. Nous tenterons de comprendre l'émergence et les développements de la cursive diplomatique dans ces documents, ce qui nous permettra de mieux saisir par ailleurs ce genre très particulier de la quittance, encore bien mal connu en diplomatique médiévale. L'exposé sera assorti de réflexions méthodologiques autour de cette enquête, associant *digital humanities*, paléographie et diplomatique.

Jinna Smit
(University of Amsterdam/Nationaal Archief)

Automatic writer identification: the paleographer's new best friend?

In the digital age, gaining insight into organization and functioning of a government administration is pretty straightforward: websites offer organization charts, procedures and guidelines. Officers can be contacted by phone or email; their pictures even show what they look like. However, to describe the chancery of the counts of Holland during the period 1299 until 1345, one can only turn to the documents produced by this institution. Although not all records were retained, their amount gives an indication of the production rate while the variety of documents provides clues about the range of tasks. The producers of the records remain unknown though, as they are rarely mentioned. Therefore, to learn more about these chancery clerks, another strategy is called for. In this case, the expression "verba volant, scripta manent" certainly rings true. We might not know the names of the writers of the chancery documents, but we can identify them by their handwritings. This makes it possible to follow their activities, responsibilities and career paths within the chancery. In the field of chancery studies, the paleographical approach is considered as tried and tested, but isn't this wishful thinking? After all, such conclusions are based on a method which increasingly has been called "ambiguous", "subjective", "authoritarian" even. Attempts to make their discipline more scientific have led to paleographers favoring measurements above observations. This not only makes it easier to describe the similarity or dissimilarity between handwritings, identifications also become more objective and verifiable. However, the quantitative approach has its limitations as well. A recent development is the use of computers for writer identification. In my own quest for a more scientific procedure, I experimented with the Groningen Intelligent Writer Identification System (GIWIS) which was developed by the Rijksuniversiteit Groningen. Far from being the first automatic writer identification system, its unique features make GIWIS highly suitable to be applied on medieval documents. During this presentation I will describe the techniques employed by GIWIS, the required input of the researcher, my test results, their reliability and future research directions. The goal of this paper, however, is not to promote a particular system, but to work towards a more systematic and extensive investigation into the (im)possibilities of digital paleography.

Dominique Stutzmann
(Institut de recherche et d'histoire des textes, Paris – FR)

Diplomatique et paléographie numerique: De nouveux instruments de datation et d'attribution

Les développements méthodologiques aussi bien que technologiques des paléographes, généralement attachés aux écritures livresques (non datées, non localisées, autographie) sont applicables à la diplomatique. Les humanités numériques en paléographie apportent ainsi des instruments nouveaux pour répondre à des questions traditionnelles : attribution (auteur, impétrant, faussaire) et datation (relative). Elles permettent aussi d'approcher les évolutions des pratiques graphiques au sein de chancelleries et de *scriptoria*.

L'examen paléographique de la production diplomatique des évêques d'Autun et de Langres, ainsi que celle des ducs de Bourgogne au XII^{e} siècle permet de mesurer exactement les apports des nouvelles méthodes pour la diplomatique. Le chartrier de Fontenay (O.Cist, dioc. Autun) est soumis à une exploration inédite pour distinguer la production de chancellerie et par l'impétrant, et les conclusions sont élargies et confrontées aux chartes des anciennes abbayes de Clairvaux (dioc. Langres), La Bussière (dioc. Autun) et Molesme (O.S.B., dioc. Langres). L'analyse des systèmes graphiques des scribes (emplois des allographes, utilisation des abréviations, capitalisation) et de la mise en texte des chartes permet de distinguer des groupes (clustering) et des évolutions dans des productions apparemment homogènes, et de séparer les productions livresque et diplomatique. Elle permet ainsi de proposer des dates pour des chartes non datées et de confirmer l'attribution, mais aussi de voir les tendances conservatrices de *scriptoria*, qui se distinguent des pratiques des chancelleries émettrices, même quand celles-ci sont bien attestées. Cette analyse ne dispense néanmoins pas des méthodes actuelles de la critique historique ou paléographique.

Johan Åhlfeldt

(http://www.francia.ahlfeldt.se, Stockholm – SE)

The Regnum Francorum Online historical GIS: a geospatial and temporal gateway to online sources of Late Antiquity and Early Medieval Europe

The Regnum Francorum Online (RFO) geographical information system explors new possibilities in digital humanities and in particular, medieval history and diplomatics, as a response to, and taking advantage of the growing number of digitized sources available online, and especially, the possibility to reference even individual documents within source-editons, using persistant identifiers (URI). Source documents in this context, refer primarily to charters, however, narrative sources, coins and inscriptions are also considered for inclusion in the database. The database extends into Late Antiquity to give the necessary historical background to medieval civilization. The long term goal of the database project is to reference (all) source-documents of Merovingian and Carolingian Europe, regarding properties of the documents themselves (authencity, origin, edition) and evidence of historical events, found within them. Properties of events, such as time, location, agent and aim are compiled within the database, together with evidence of medieval entities such as place, institution, territory, personal names and person, mentioned in the events. The database application renders two-way interactive maps based on the meta-data of documents and events, respectively, together with background layers of medieval territories, roman roads, modern extension of forest, and other features relevant to historical analysis. This makes it possible to visualize on maps the distribution of documents with different characteristics, and the itinerary and activities of rulers, development of posessions of institution, extension and change of territories, in time and space. Meta-data can also be visualized thematically, e.g. the distribition of fiscal property, of churches and monasteries, etc. RFO contains historical information about 14,000 places, 1,500 institutions, 600 medieval territories, and other entities, in the realm of the Frankish kingdom/empire. This far, 18.000 source-documents have been referenced and analyzed. For the purpose of inter-linking and embedding external documents and maps into the RFO-gateway, identifiers of external source-documents are maintained within the database as they become available on the internet. This way, it serves as a geographical gateway to scattered online sources.

Redmer Alma

(Drents Archief, Assen – NL)

The "Digitaal Oorkondeboek Groningen en Drenthe" and the Semantic Web:towards ageneric publication of Dutch medieval sources?

In 2007, the first phase of the project DOGD (Digital Charterbook Groningen and Drenthe) was completed and the website www.cartago.nl was launched. As a result of this project more than 20.000 charters and other sources from the provincial archives in Groningen and Drenthe were that year made digitally accessible, recently extended to 150.000 images.

The early charters and deeds in the Dutch provinces Groningen and Drenthe are the most important historical sources of these regions. There are hardly any older city accounts, judicial protocols or other comparable sources that can be systematically researched up to the mid-16thcentury. This explains the importance of charters in these Dutch provinces, sharply contrasting with the difficulty in retrieval of these sparse but disparate extant sources.

The project's website www.cartago.nl is a newly developed system that makes all relevant documents accessible with scans of original documents, existing metadata and transcriptions that are also searchable.

Today the DOGD is the only larger digital charterbook in the Netherlands. Various archives in the Netherlands have shown interest in the possibility to publish different medieval sources (chartes, city accounts etc.), using the technology and the experiences of the website Cartago.

That enabled the DOGD to take the lead to design a more generic structure to meet the needs of archivists and users, scientists as well as non-professionals. A large, general publication of all Dutch medieval sources will never be possible (and probably not desirable). The Semantic Web however offers possiblities to connect current and future digital charterbooks, can help to prevent re-inventing the wheel and to work towards a next generation of publication of (Dutch) medieval sources in a European context.

Aleksandrs Ivanovs, Aleksey Varfolomeyev

(Daugavpils University – LV/ Petrozavodsk State University – RU)

Semantic Publications of Charter Corpora (The Case of a Diplomatic Edition of the Complex of Old Russian Charters ‘Moscowitica–Ruthenica’)

Nowadays, the term ‘semantic publication’ is generally accepted. It is commonly used to denote an electronic text publication that is provided with additional information layer, which represents the sense of the text (i.e. knowledge about the text) in a formalized way suitable for automatic processing. In the modern Web environments, semantic publications, especially in digital libraries and electronic journals, have become quite topical. The advantages of semantic publications are as follows: firstly, they provide better facilities for searching for information. Secondly, such publications can be used as knowledge bases in order to generate (by means of automatic inference) new knowledge and/or hypotheses for further research. In this regard, potential advantages of semantic editions of historical records are rather obvious (see Ahonen and Hyvönen 2009; Mirzaee, Iverson and Hamidzadeh 2005). However, generally approved conceptual and technological approaches to designing semantic publications have not been elaborated yet.

This paper shows how the basic principles of semantic publications can be applied to electronic scholarly editing of charter corpora. In order to reveal advantages of such editions in diplomatic and historical research of medieval charters, the authors present a multifunctional prototype of a semantic publication of the 13th century Old Russian charter corpus – a constituent part of the vast collection of medieval and early modern records ‘Moscowitica–Ruthenica’ kept in the Latvian State Historical Archives (Riga). These documents describe relations of Old Russian and Byelorussian lands and towns with Riga, Livonia, and Hanseatic League in the late 12th – early 17th centuries. The prototype of the semantic publication is designed as a comprehensive diplomatic edition of Old Russian Charters that represents paleographic features of the documents. In the diplomatic transcription, well-developed TEI and CEI markup schemes and elements can be used. At the same time, semantic nature of the publication poses a number of problems. E.g., with some exceptions, Old Russian symbols can be transcribed using Unicode symbols, however, sometimes it can come into conflict with the present-day semantics of the symbols: in the 12th–13th centuries, the graphic form of the Cyrillic letter ‘*И*’(‘i’) was similar to the graphic form of the modern Cyrillic letter ‘Н’(‘n’); Old Russian ‘Н’(‘n’) usually resembled Latin ‘N’, etc. In the diplomatic transcription of Old Russian charters, original graphic forms of the letters should be comprehensively reproduced in order to create appropriate tools for paleographic dating and attribution of the documents. Unfortunately, this mode of representation of Old Russian symbols may hinder analytical and searching operations. The paper proposes a number of solutions to the problem: automatic substitution of symbols in the texts that appear on display, elaboration of special fonts on the basis of SVG technologies, etc. In order to provide the texts with additional data (descriptive metadata), information about persons, sites, events, etc. mentioned in the charters is revealed and linked with the corresponding data extracted from different specialized ontologies. In the last years, a great number of different ontologies have been created, including those intended for historical and source studies, e.g. ontology CIDOC CRM for the purposes of description of museum objects. On the basis of this ontology, specialized historical ontologies have been developed (e.g. see Ide and Woolner 2007). In contrast with the event-oriented approach accepted in the above-mentioned ontologies, the authors of the paper propose a document-oriented approach to description of historical data. However, production of semantic publications on the basis of ontologies, which are recorded using Semantic Web technologies – RDF or OWL, is time-consuming. It seems that opportunities and tools provided by semantic Wiki-systems can facilitate this process. For instance, Semantic MediaWiki offers special, rather simple markup tools that can be used to indicate different objects (place-names, persons' names, etc.) in the texts of the charters and to supply the texts with meta-information. Since the semantic edition is based on a detailed markup of the texts, it can also provide appropriate tools for an in-depth pattern analysis of the charters. The structure of the texts of Old Russian charters is rather definite and hierarchical and, on the whole, does not differ much from that of European medieval charters. Thus, the texts of the documents can be usually divided into such semantic parts as protocol, (main) text, and eschatocol; within these parts smaller fragments can be marked out, e.g. invocation, intitulatio, inscriptio, salutatio, etc. in the charters' protocols. At the same time, semantic designations of structural parts (articles, chapters) and clauses that constitute main texts can not be normalized completely. Therefore, while some of the above-mentioned semantic parts of the texts can be tagged using CEI and TEI elements, other elements should be introduced anew to meet the requirements of this semantic publication. In order to represent (e.g. in English) the sense of the charters' structural parts and smaller fragments, controlled natural languages like Attempto Controlled English (ACE) can be used. Texts in ACE can be automatically translated into First Order Logic formulae, thus enabling logical inference, acquiring new knowledge, and creation of specific models of historical reality. The semantic publication can be used as a basis for a specific Web information system that incorporates charters' texts, research tools, and research results into a knowledge-based system, which is created for a network community.

References

Ahonen, E. and E. Hyvönen (2009) "Publishing Historical Texts on the Semantic Web – A Case Study," in: *Proceedings of the Third IEEE International Conference on Semantic Computing (ICSC2009)*. Berkeley. Pp.167-173.

Ide, N. and D. Woolner (2007) "Historical Ontologies," in: K. Ahmad, C. Brewster, M. Stevenson (eds.) *Words and Intelligence II: Essays in Honor of Yorick Wilks*. [S.l.]: Springer. Pp.137-152.

Mirzaee, V., Iverson, L. and B. Hamidzadeh (2005) "Computational Representation of Semantics in Historical Documents," in: Humanities, Computers and Cultural Heritage: Proceedings of the XVIth International Conference of the Association for History and Computing. Amsterdam. Pp.199-206.

Publications sémantiques de corpus des chartes (le cas d'une édition diplomatique du complexe des anciennes chartes russes 'Moscowitica–Ruthenica')

Publication sémantique de Corpora des Chartes (Exemple 'Moscowitica–Ruthenica')

Aujourd'hui, le terme 'publication sémantique' est généralement admis. Il est couramment utilisé pour désigner une publication du texte électronique qui est fourni d'une couche d'information supplémentaire, ce qui représente le sens du texte (ou les connaissances sur le texte) d'une manière formalisée adapté pour le traitement automatique. Dans les environnements Web modernes, publications sémantiques, en particulier dans des bibliothèques numériques et journaux électroniques, ont devenus très actuelles. Les avantages de publications sémantiques sont comme suit: premièrement, ils fournissent des meilleures possibilités pour la recherche des informations. Deuxièmement, ces publications peuvent être utilisées comme bases de connaissances, ce qui peut générer de nouvelles connaissances et/ou des hypothèses pour des recherches suivantes au moyen de l'inférence automatique. À cet égard, les avantages potentiels des publications sémantiques des documents historiques sont assez évidents (Ahonen and Hyvönen 2009; Mirzaee, Iverson and Hamidzadeh 2005). Toutefois, les approches conceptuelles et technologiques, qui sont généralement approuvé pour la conception de publications sémantiques, ne sont pas élaborées encore. Cet article montre comment les principes de base de publications sémantiques peuvent être appliqués à l'édition savante électronique d'un corpus des chartes. Afin de révéler les avantages de ces éditions dans la recherche diplomatique et historique de chartes médiévales, les auteurs présentent un prototype multifonctionnel d'une publication sémantique du corpus des anciennes chartes russes du XIII siècle – une partie constitutive de la vaste collection de documents historiques médiévales et modernes 'Moscowitica–Ruthenica' qui est conservé dans les Archives Historiques d'Etat de Lettonie (Riga). Ces documents décrivent les relations des terres et des villes russes et biélorusses de Riga, de Livonie, et de la Hanse, à la fin du XII – début du XVII siècles. Le prototype de la publication sémantique est conçu comme une édition diplomatique globale du corpus des chartes qui représente les caractéristiques paléographiques des documents. Dans la transcription diplomatique, les schémas bien développés de balisage TEI et CEI peuvent être utilisés. Dans le même temps, la nature sémantique de la publication pose un certain nombre de problèmes. Par exemple, en dehors de quelques exceptions, les vieux symboles russes peuvent être transcrits en utilisant des symboles Unicode. Parfois, la transcription peut entrer en conflit avec la sémantique actuelle des symboles: dans les XII – XIII siècles, la forme graphique de la lettre cyrillique 'И' ('i') était semblable à la forme graphique de la lettre cyrillique moderne 'Н' ('n'); la vieille lettre russe 'Н' ('n') en général ressemblait à 'N' latine, etc. Dans la transcription diplomatique, les primaires formes graphiques des lettres doivent être complètement reproduite dans le but de créer des outils appropriés pour la datation paléographique et l'attribution des documents. Malheureusement, ce mode de représentation des vieux symboles russes peut entraver les opérations d'analyse et de recherche. L'article propose un certain nombre de solutions au problème: la substitution automatique des symboles dans les textes qui apparaissent sur l'écran, l'élaboration de polices de caractères spéciaux sur la base des technologies de SVG, etc. Afin de fournir les textes avec des informations supplémentaires (métadonnées descriptives), des informations sur les personnes, de lieux, événements, etc. mentionnés dans les chartes sont révélées et liées aux données correspondantes extraites de différentes ontologies spécialisées. Dans les dernières années, un grand nombre de différentes ontologies ont été créés, y compris ceux pour l'histoire et les études des sources, par exemple, CIDOC CRM ontologie pour description des objets de musée. Sur la base de cette ontologie, spécialisée ontologies historiques ont été développés (Ide and Woolner 2007). En contraste avec l'approche événementielle accepté dans les ontologies mentionnées ci-dessus, les auteurs de l'article proposent une approche à la description des données historiques centrée sur les documents. Cependant, la production de publications sémantiques sur la base d'ontologies, qui sont enregistrés en utilisant les technologies du Web sémantique – RDF ou OWL, prend beaucoup de temps. Il semble que les possibilités et les outils fournis par Wiki sémantique peuvent faciliter ce processus. Par exemple, Semantic MediaWiki offres des outils de balisage simplifiés qui peut être utilisé pour indiquer les différents objets (noms de lieux, noms de personnes, etc.) dans les textes des chartes et de fournir les textes avec des méta- informations. Depuis la publication sémantique est basée sur un balisage détaillé de textes, elle peut également fournir des outils appropriés pour une analyse profonde des structures des chartes. La structure des textes des anciennes chartes russes est définie et hiérarchisée et, dans l'ensemble, ne diffère pas beaucoup de celle des chartes européennes médiévales. Ainsi, les textes peuvent être généralement divisés en parties sémantiques telles que le protocole, le texte, et l'eschatocole; dans ces parties les fragments plus petits peuvent être balisés, par exemple, invocation, suscription, adresse, salutation, etc. dans le protocole de charte. En même temps, les désignations sémantiques de pièces de structure et des clauses, qui constituent le texte principal, ne peuvent pas être normalisées entièrement. Par conséquent, si certaines pièces sémantiques des textes peuvent être étiquetées en utilisant des éléments des schémas CEI et TEI, d'autres éléments doivent être introduits de nouveau en respectant les exigences de cette publication sémantique. Afin de représenter (par exemple en anglais) le sens des pièces de structure et de petits fragments de la charte dans la publication sémantique, les langues contrôlées naturelles comme Attempto Controlled English (ACE) peuvent être utilisées. Les textes en ACE peuvent être automatiquement traduit en formulas de la logique des prédicats du premier ordre, permettant ainsi d'inférence logique de nouvelles connaissances, et la création de modèles spécifiques de la réalité historique. La publication sémantique peut être utilisée comme une base pour un système d'information Web spécifique qui intègre les textes des chartes, les outils de recherche, et les résultats de recherche dans un système basé sur les connaissances, qui est créé pour une communauté scientifique en ligne.

Les references

Ahonen, E. and E. Hyvönen (2009) "Publishing Historical Texts on the Semantic Web – A Case Study," in: Proceedings of the Third IEEE International Conference on Semantic Computing (ICSC2009). Berkeley. Pp.167-173.

Ide, N. and D. Woolner (2007) "Historical Ontologies," in: K. Ahmad, C. Brewster, M. Stevenson (eds.) Words and Intelligence II: Essays in Honor of Yorick Wilks. [S.l.]: Springer. Pp.137-152.

Mirzaee, V., Iverson, L. and B. Hamidzadeh (2005) "Computational Representation of Semantics in Historical Documents," in: Humanities, Computers and Cultural Heritage: Proceedings of the XVIth International Conference of the Association for History and Computing. Amsterdam. Pp.199-206.

Michael Hänchen M.A.

(Research Center for Comparative History of Religious Orders TU Dresden)

Development, formalization and capacity of evaluation of digital charters by means of adatabase

1.) The Project

Using the capacity of modern information technology, the development of charters experiences a fundamental change. The digital photography of great collections of charters enables an immediate research thanks to using high resolution pictures as in the virtual research environment www.monasterium.net. My project at the Research Center for the Comparative History of Religious Orders at the University of Dresden neither assembles nor makes available new source material; rather, it analyses with regards to content the charters already available on Monasterium.net under the historical perspective of documentarily bequeathed memorial donations in the 14th century. In this period, which is called the "century of crisis", significant ecclesial, economical and socio - cultural events occurred and characterized the expectations of the last days. The possible impacts of these events endow a comparative study of the behavior of the benefactors with the giving of memory's gifts. The main goal of my database - assisted study is the investigation of the mentality of donation - giving by different social groups in the southern German region in view of these events. The referred social groups are the aristocracy, the lower nobility, the citizens and the clergy. Two aspects have priority within this study. The first question comes in the "textualisation of law" (Verschriftlichung von Recht) in the memorial charters. Are there recognizable processes of change in the context of the historical events and are there regional or social differences or changes in the charters' outer and inner features? Secondarily the study researches the progress of the numbers of donations, the social groups and the persons, the material aspects of the donations, the receivers of these donations and the spiritual reward expected by the benefactors. According to the comparative perspective, the study discusses the question, which relations existed between religious, juridical and economical factors in the commemoration during the 14th century? This means preferential treatment of particular orders and monasteries by different social groups, which acts of donations - and which forms of donations are noted in charters and what was the economical potency of the benefactor and the recipients of gifts. Furthermore, I will present you the system I developed for the registration and formalization of the charters. Thereafter I will explain the handling of the database. Lastly I will illustrate the initial results with the help of short examples.

2.) Development of digital charters by means of a database With the help of a database it's possible to concentrate digital charters for further specified evaluations. That for I reckon, read and carry the digital images in a database matrix. This means that every charter corresponds with one data record. The matrix is structured in 11 categories, according to the aforementioned perspectives of perception and registers - formal (I. - IV.) and contextual (V. - XI.) aspects of the images as consultable data sets. Category I. (F1) gathers the meta data of charters, Category II. (F2) the outer features. Category III. (F3) records the inner features of the charter. As last formal category, point VI. (F4) gathers 2 the important piece of information about the ecclesial recipient of the donation. In Category . (F5) ‐ the art of donation ‐ the matrix collects contextual charter's data's. Category VI. Specifies the benefactor and the intention of the donation. Category VII. (F6) defines initially the material aspect of the donation which is later specified by point VIII. I use keywords so that it's possible to query the database for the exact type of the donation. Next up point IX. (F7) records the spiritual reward. And point X. (F8) save the clauses by the benefactor compared to the donated convent and point XI. which handles the clauses by the convent compared to the benefactor.

3.) Formalization of digital charters by means of a database With the help of this matrix it's intended to formalize and to normalize the images as much as possible in order to query the database under similar terminological conditions. The principal goal is to describe both the charters with the terms of diplomatics (*method of affixing the seals*; pushed through/impressed – *direction of the format*; Charta diplomatica/charta transversa) and to antagonize the variances of writing in the charters funds. These variances in the charter reflect often one and the same name and term but in different ways. For all places, monasteries and dominions, insofar as allocable, I use the modern, official and customary terms (*Pazzaw*/Passau). Also the names of persons (*Hainrich, Hinrich*/Heinrich), dynasties (*Leubolfing*/Leiblfing), the social rank and titles (*miles*/Knight), material of the donation (*hube*/Hufe) and relationship (*Vorvadern*/Vorfahren) will be registred under modern terms. For established abbreviations (*Cistercian Order*/ OCist) I use the Lexikon of Theologie und Kirche (LThK). The type of dating by the ecclesial feasts and saints is also formalized (*sand giligen tag* /S. Ägidius), whereby we have to consider the regional characteristics. But the real date is less important than the connection to a saint at the moment of the donation. I am aware that this formalization includes difficult border cases as well. For instance formalization into modern terms cannot be done always in the same manner. For these ambiguous cases the database gives the option to record variances and additional information. These could be fixed if new information is available.

4.) Evaluation competences of the digital charters with the help of a database Using a database offers the possibility to find the relevant criteria of the study very quickly. Its evaluation can be done with singular and multiple terms scanning. Every combination of search terms is conceivable and possible. If, for example, the user is looking for the format, the language or the font of a charter at a certain time or after a certain group within a determinate territory, the database will find the reliable charters. To illustrate this, I want to present some examples based on two Cistercian monasteries, Aldersbach and its filiation Fürstenzell. Both are located in an agrarian region near the city of Passau (F9). The potential of the database for these two monasteries is great because their charter funds are without recognizable gaps. There is a total of 490 charters from Aldersbach and 369 charters from Fürstenzell during the period of 1300 ‑ 1400. Within 77 donations in memory of Aldersbach and 50 from Fürstenzell are conserved. Added together we have a total of 127 pieces.

As for the "textualisation of law", it can be argued that the charters are the fully subjective written charter of the late Middle Age (127/100%). The predominant language is German. In these 127 charters 121 are in German and 6 in Latin. The Latin charters are invariably made out by clerks. For instance 3 charters of the churchmen clarify that the German language has entered the "textualisation of law". When it comes to the size of the charters we can see 3 that the most commonly used size of the parchment is octavo (86 pieces/67%). 15 charters (12%) are in quarto size, 18 (14%) are mixed formats of octavo and quarto and finally 8 pieces (6%) are in folio or larger. A uniform format is not recognizable: Many more of the charters are close in size to the aforementioned formats. It appears in both monasteries, that here German almost completely replaced Latin in the memorial documents between lay people and clerics and that the oft ‑ used octave format can be evaluated as a pragmatic consequence of the ever ‑ increasing written records. Through the comparative citation of further charter material it can be demonstrated whether it involves regional, order or monastery ‑ specific developments in the Passau area or whether general tendencies of the "textualisation of law" already present themselves here in the 14th century. An interesting tendency is shown in the number of donations (F10). 5 Years before the plague strikes the region of Passau (in 1349) the numbers of donations decrease significantly at both Cistercian monasteries. Up to 1400 there is no considerable increase to attest. The plague itself cannot be the reason for this decrease of the donations. There are likely other reasons to consider: for example an economical downfall of the region, a loss in attractiveness of the Cistercian order or regional political events. However because the desire for liturgical commemoration of the medieval people was essential, the question arises to which other monasteries, orders or institutions the donations were going at that very time. Another tendency shows the "requirements of the liturgical commemoration" (F11). Could it be possible that the terrible memories of the plague had lead to an increase in "explicit requirements to liturgical commemoration" at the time immediately subsequent? On the contrary, the charters show that the "requirements" to Aldersbach and Fürstenzell remained as numerous as before the plague. Finally, I want to discuss the benefactor and the material of the donation in terms of the Category VIII. "Benefication" (F12). While Aldersbach (38 pieces) acquired 25 permanent monetary receipts (67%) Fürstenzell got (altogether 22 pieces) with 10 pieces (45%) less. But

Fürstenzell became with 6 charters (27%), more donations of tithe than Aldersbach with only 1 (3%). Only Aldersbach received 4 finacial donations (11%). The number of donations in natural products is at both monasteries nearly equal (Aldersbach 7 [19%], Fürstenzell 5 [23%]). And a mixed donation with natural products and permanent money earnings is only findable at Fürstenzell with one Charter (5%). The predominant monetary donations during the 14th century, especially at Aldersbach, could support the scientific paradigm that the Cistercian Order changed their established type of economical subsistence strategy from only own management to an economical strategy based on earnings at the end of the 13th century.

The benefactors of the afore - mentioned donations (F13) are: Aldersbach got donated by 31 (81%) laymen (Knights, lower nobility etc.), 5 (13%) citizens (one from Straubing and Passau, one from Krems in the duchy of Austria, 2 from the proximate city of Vilshofen), one from the count of Ortenburg and one by the rival king Friedrich III. Fürstenzell got donated by 16 (73%) laymen and by 6 (27%) citizens from Passau but without any donation from the high nobility as we can find in Aldersbach. It seems then that the Cistercian Order is preferred by the landed gentry. The counts of Ortenburg are the holders of the bailiwick of Aldersbach.

To sum up: which information could be extracted in context of the aforementioned questions? It has to be considered, however, that the examples presented today are still very selective and limited in their significance on a local context. About the first question about "textualisation of law" we can say that all charters are the fully developed charters of the late Middle Ages with Intitulatio, Publicatio, Dispositio, Corroboratio and Datatio. The predominant format is octavo and concerning the language, German is preferred with few variations of the inner and outer features in context of the social groups. About the second question we can say that there was a significant decrease of donations before the plague and we have differentiated the "type of donations" and the "benefactors". Aldersbach got donated with a lot more permanent and singular benefits in money than Fürstenzell, whichreceived more donations in tithe. Nearly equal are the numbers of natural product donations in both monasteries.

We also distinguish differences among the benefactors. Aldersbach was donated by the lower and higher nobility and by citizens from cities like aforementioned. The Cistercian monastery of Fürstenzell was only donated by the lower nobility and citizens only from the nearby diocesan city of Passau. A change or a rise of the "explicit spiritual rewards" by the benefactors in the charters could not be proved yet. With the help of a great number of charters, using and evaluating a database can show tendencies and developments along a defined question and it is a useful way to open up digital corpora for further research. Furthermore linked with other sources and other scientific literature, the usage of a database can open a window to cultural and prosopographic history of donations in memory by means of the charters.

Martin Roland

(Österreichische Akademie der Wissenschaften Kommission für Schrift- und Buchwesen des Mittelalters, Wien – AT)

Illuminated Charters in the Digital Age – Rules and Opportunities

Illuminated charters, i. e. charters with historiated decoration or additional colours, were picked out as special examples of diplomatics as early as 19th century. A considerable number of studies were published but a general overview is lacking and there is almost no *Codex diplomaticus* or respective database systematically mentioning drawn or painted decoration. A targeted search for illuminated charters is impossible, neither in an analogue nor a digital way.

As a starting point I present a number of illuminated charters to focus on the pan-European dimension of the phenomenon. The cultural historical importance of interesting and individual items is unquestioned; the dowry charter of Theophanu was even proposed to become UNESCO-World Heritage. On the other hand there are groups just as collective indulgences or „Wappenbriefe“ (granting coats of arms) of which even today hundreds of examples have survived.

Which new opportunities arise from the digital age? *Monasterium.net* offers more than 200.000 digitized charters on the web; many archives have started comparable projects. However, without proper metadata the treasures hidden in these digitized holdings cannot be discovered. As a first step, that is very easy to be implemented, the existing online-descriptions could be accompanied by thumbnails. Thus a first glimpse of the external features would be possible.

To use the possibilities of the digital age to their full potential some rules must be observed. At first the object of interest must be defined properly: My paper proposes a definition of what an illuminated charter exactly is. The definition mentioned above – historiated and/or additional colours – is valid for, presumably, not more than 1000–1500 charters until 1520. Therefore I propose to define two additional levels of illuminated charters: These comprehend drawn decoration, very often characteristic for specific chancelleries (level 2) and graphic means of authenticating (level 3). Furthermore, I present a list (in German, English, Italian and French) of 15 key words (Normbegriffen) which will facilitate a comprehensive search. The description of the charter and its decoration will be tagged with the respective terms thus allowing search within a database but also via conventional search engines. The potential of the digital age is the vast availability of digital images of charters, the rules to be observed are deliberate categories and consistent terminology (key words) to facilitate well targeted search.

Illuminierte Urkunden im digitalen Zeitalter – Maßregeln und Chancen

Illuminierte Urkunden, das heißt Urkunden, die historisierten Schmuck oder mit Farbe aufgetragene Elemente enthalten, wurden schon im 19. Jahrhundert als Sonderfälle der Urkundenpraxis thematisiert. Es gibt zwar zahlreiche Einzelstudien aber keine zusammenfassende Darstellung und kaum ein Urkundenbuch oder eine entsprechende Datenbank verzeichnet systematisch den eventuell vorhandenen gemalten oder gezeichneten Schmuck. Das gezielte Suchen illuminierter Urkunden ist daher derzeit – weder analog noch digital – möglich.

In einem ersten Schritt werden Beispiele illuminierter Urkunden vorgestellt und deren gesamteuropäische Dimension beleuchtet. Neben interessanten Einzelstücken, deren kulturhistorische Bedeutung unbestritten ist, die Dotalurkunde der Theophanu wurde sogar als Weltkulturerbe vorgeschlagen, gibt es Gruppen wie kuriale Sammelablässe oder Wappenbriefe, von denen auch heute noch hunderte Stücke erhalten sind.

Welche Chancen bietet nun das digitale Zeitalter? Monasterium.net stellt über 200.000 Urkunden im Netz zur Verfügung, viele Archive bieten ebenfalls zahllose Digitalisate an. Doch ohne entsprechende Metadaten bleiben die enthaltenen Schätze an illuminierten Urkunden trotz Digitalisierung weiterhin verborgen. Als erster, ganz einfach zu verwirklichender Schritt könnte neben den Beschreibungen (beim Browsen und bei den Treffern) auch Thumbnails der Digitalisate vorgesehen werden, die dem Suchenden einen ersten Eindruck der äußeren Merkmale der Stücke vermitteln.

Um die Chancen des digitalen Angebots aber umfassend nutzen zu können, müssen gewisse Maßregeln eingehalten werden. Zuerst muß der Untersuchungsgegenstand definiert werden: Im Rahmen dieses Vortrages stelle ich eine Definitionen für den Begriff „illuminierte Urkunde“ vor. Die oben genannte Definition (historisiert und/oder zusätzliche Farbe) trifft nur auf vergleichsweise wenige Beispiele zu (vielleicht 1000–1500 Urkunden bis ca. 1520). Es wird daher vorgeschlagen, zwei zusätzliche Niveaus von illuminierten Urkunden zu definieren: Diese umfassen auch gezeichneten Schmuck, der sehr oft für einzelne Kanzleien typisch ist (Niveau 2), sowie graphische Beglaubigungszeichen (Niveau 3).

Weiters stellen wir eine Liste (in deutsch, englisch, italienisch und französisch) von 15 Normbegriffen (standardisierte Suchbegriffe) vor, mit denen illuminierte Urkunden suchbar gemacht werden können. Die weiterhin frei formulierten Urkundenbeschreibungen werden mit diesen

Normbegriffen getaggt und ermöglichen so die gezielte Suche sowohl innerhalb des Systems als mittels Suchmaschinen (Google).

Die Chance des digitalen Zeitalters ist die massenweise Verfügbarkeit von Abbildungen, zu befolgende Maßregeln sind durchdachte Kategorien und einheitliche Normbegriffe für die gezielte Suche.

Žarko Vujošević

(Institute for Balkan Studies of the Serbian Academy of Sciences and Arts)

The Medieval Serbian Chancery: Challenge of Digital Diplomatics

At the time when digital information technologies are increasingly entering the field of processing and presentation of documentary heritage, Serbian diplomatics still hasn't published its comparatively modest corpus of less than a thousand medieval documents within one all-encompassing collection. The question therefore arises as to whether it would be better to switch the efforts aimed at producing a traditional printed edition towards producing a digital one. Simple, quick and reliable access to digitized data would considerably enhance the use of diplomatic sources in research of almost all aspects of medieval Serbian past. In addition, uploading of digital contents to the internet should inspire wider interest in Serbian documentary material as a distinctive blend of three diplomatic traditions (Slavic, Latin and Byzantine), also reflected in the use of three diplomatic languages, and to facilitate its study in a European context. Most importantly for Serbian diplomatics as a scholarly discipline, this *born digital* project will give strong encouragement to dealing with the inadequately studied key issues relating to the functioning of the chancery, that is, the process of document creation. Fitted into digital databases, the ample and diverse information about the creators of Serbian charters, the conditions in which they worked and the patterns they might have followed, would be a vital prerequisite to a comprehensive inductive study of the actual degree of bureaucratic formalization of Serbian documents. Through this, it should be possible to establish whether they were indeed products of a system, represented by an institution commonly termed 'the chancery', or improvisation and *ad hoc* solutions, typical of medieval society.

Anaïs Wion

(CNRS, Centre d'Études des Mondes Africains, Paris)

Editing vs. Analysing? The Intellectual Cost of Digital Humanities through the example of the Ethiopian manuscript Archives project

What is meant by "archives" in the Ethiopian context?

This global term includes administrative, legal and historical texts, which were produced by the Ethiopian political and religious powers to witness their laws, their rules, their traditions. The producers of these documents, between the thirteenth and the twentieth century, are the royal administration and to a lesser extent religious administrations. They are written in Ge'ez, the ritualised language of the Ethiopian Christian culture, a Semitic language with its own writing system, an alpha-syllabary. Private acts are coming later, from the mid-eighteenth century onwards and they are written in Amharic, the vernacular language of the Ethiopian highlands, written with the same writing system as his indirect ancestor Ge'ez.

These documents are copied, in most cases in the blanks pages of the biblical or liturgical manuscripts. Thus the margins of Gospels manuscripts, or Miracles of Mary contain the archival documents of the kingdom. This dispersion of archives in the books of the monastic libraries is one of the reasons for the lack of consideration given to them by researchers until now. A few thousand, maybe even hundreds of thousands of documents of various kinds are a result of corpus sources largely untapped to date.

The EMA project is one step in the construction of an Ethiopian diplomatic.

This presentation is at the same time retrospective, analyzing the maturation of this project since 2005, analytical, describing the current state of work and prospective in addressing the research community gathered at this conference a series of questions on the future development of Diplomatics and analysis of documents within the Digital Humanities.

Indeed, the choice of a structure in XML / TEI has solved some issues, as much as it has closed some doors to the wishes of the original project. Encoding is indeed a cognitive act that allows a rich reading and a very accomplished understanding. It is therefore in itself a form of analysis. However, is editing as unique achievement of this work a satisfactory output? The cost to enter the Digital Humanities, for a researcher who is not initially trained in its various technologies, is it up to what he or she can expect to gain in terms of textual and historical analysis?

Éditer vs. Analyser? Le coût intellectuel des Humanités Numériques au prisme du projet Ethiopian manuscript Archives

Qu'appelle-t-on les archives manuscrites éthiopiennes ? Il s'agit d'un qualificatif général englobant les textes de nature administrative, juridique et historique qui furent produits par les pouvoirs politiques et religieux éthiopiens pour témoigner de leurs lois, de leurs règles, de leurs traditions : écrits légaux et pseudo-légaux, documents historiographiques ou contractuels, chartes de donation de terres, règles concernant les nominations de dignitaires religieux ou fonctionnaires royaux, etc. Les producteurs de ces documents, entre le treizième et le vingtième siècle, sont l'administration royale et dans une moindre mesure les administrations religieuses. Ils sont rédigés en ge'ez, la langue ritualisée de l'Éthiopie chrétienne, une langue sémitique qui possède son propre système d'écriture, un alpha-syllabaire. Les actes privés ne sont issus que tardivement, à partir de la mi-xviii^{e} siècle et ils sont eux rédigés en amharique, la langue véhiculaire des hauts-plateaux éthiopiens, écrite avec le même système d'écriture que le ge'ez, son ancêtre indirect.

L'une des particularités de la conservation de ces documents en Éthiopie est qu'ils sont copiés, dans la majeure partie des cas, dans les espaces blancs laissés dans les manuscrits bibliques ou liturgiques. Ainsi les manuscrits des Évangiles, des Miracles de Marie contiennent dans leurs marges et entre leurs chapitres, les documents d'archive du royaume. Cette dispersion des archives au sein des livres des bibliothèques monastiques est l'une des raisons du peu de considération qui leur fut accordé par les chercheurs jusqu'à aujourd'hui. Quelques milliers, peut-être même centaines de milliers, de documents de nature diverse constituent un corpus conséquent de sources largement sous-exploitées à ce jour.

Le projet EMA est une étape dans la construction d'une diplomatique éthiopienne. Cette présentation se veut à la fois rétrospective, analysant les phases de maturation de ce projet depuis 2005; analytique, décrivant l'état actuel du travail; et prospective, adressant à la communauté des chercheurs réunis lors de cette conférence une série de questionnements sur le futur développement de la diplomatique et de l'analyse des documents dans le cadre des Humanités Numériques. En effet, le choix d'une structuration en XML/TEI a résolu des questions tout autant qu'il a fermé des portes à certain des desiderata initiaux du projet. Certes l'encodage est un acte cognitif riche qui permet une lecture et une compréhension très aboutie. Il est donc en soi une forme d'analyse. Néanmoins, éditer comme presque unique aboutissement à ce travail conséquent est-il satisfaisant ? Le coût d'entrée dans les Humanités Numériques, pour un chercheur qui n’est pas initialement formé à ses diverses technologies, est-il à la hauteur de ce qu’il peut espérer y gagner en terme d’analyse des documents textuels?

Serena Falletta

(Università degli studi di Palermo – IT)

From paper to bit. The cartulary of Santa Maria Nuova in Monreale digital edition

The aim of the project is the presentation of the model and experimental hypertext publishing, research and communication focused on a historical corpus of documentary texts encoded by the XML syntax and a labeling scheme tailored to the specific characteristics of the documentation and the needs and classical categories of analysis critique of historical research. For the purposes of research and project work, was chosen as the subject of investigation and application a bishop's cartulary, the *Liber privilegiorum Sanctae Montis Regalis ecclesiae*, in the unpublished tradition of the fifteenth century, reported from the code Vat.Lat. 3880 currently preserved at the Vatican Library: a documentary series not numerous but homogeneous, so suitable to the character of this work aimed to get a simple but rich the correlations between text and data, in an attempt to overcome the conceptual and methodological problems founded by historians in the use of rigid DataBase, inadequate to express the complexity of qualitative data. The experiment was designed as a laboratory through which to understand more precisely the not neutral nature of the informatics and deep epistemological implications that will its implementation: beyond the traditional historical discourse, it’s able to represent simultaneous and interrelated processes, offering a communication vehicle capable of incorporating different sources in the hypertext and reconfiguring technical equipment and academic practice. We speak, of course, about an “acceptable trial”, with which to prove - enhancing the telematics and its language’s specificity - the ability to trace a code that including in a wide web hypertext, the documents produced by a specific institutional subject and their history, trying to avoid the distortion of the discipline’s identity, through the use of tools and usual classifications, to meet the demand of rigorous interpretation. It is therefore a digital history experiment, that contains and make visible the documentation’s links, reflecting the dynamism, but whose main characteristics remain flexible, plurality of paths for consultation and differentiation of possible approaches.

Dalla carta al bit. L'edizione digitale del cartulario di Santa Maria Nova di Monreale

Scopo dell'intervento è la presentazione del modello ipertestuale e sperimentale di edizione, ricerca e comunicazione storica incentrato su un *corpus* di testi documentari codificati adottando la sintassi XML e uno schema di marcatura calibrato sulle caratteristiche specifiche della documentazione e sulle classiche esigenze e categorie di analisi critica della ricerca storica. Ai fini della progetto di ricerca e lavoro, si è scelto come oggetto d'indagine e applicazione un cartulario vescovile, il *Liber privilegiorum Sanctae Montis Regalis ecclesiae,* nell'inedita tradizione quattrocentesca riportata dal codice *Vat. Lat. 3880* attualmente conservato presso la Biblioteca Apostolica Vaticana: una serie documentaria non numerosa ma omogenea, adatta quindi al carattere di questo lavoro, finalizzato ad ottenere in modo semplice ma ricco le possibili correlazioni tra testi e dati, nel tentativo di superare i problemi concettuali e metodologici riscontrati dagli storici nell'utilizzo dei rigidi *DataBase,* inadeguati ad esprimere la complessità di dati qualitativi. L'esperimento condotto è stato concepito come un laboratorio attraverso cui comprendere con più precisione la natura *non neutra* del fenomeno informatico e delle profonde implicazioni epistemologiche che la sua applicazione comporta: superando il tradizionale discorso storico, caratterizzato da linearità, esso è infatti in grado di rappresentare processi simultanei e interconnessi, proponendo un veicolo comunicativo in grado di incorporare ipertestualmente fonti disparate, e riconfigurando così strumentazione tecnica e prassi accademica. Si parla, chiaramente, di una *sperimentazione sostenibile,* con la quale esperire – valorizzando il mezzo telematico e la sua specificità di linguaggio – la possibilità di ricostruire un codice che ricomprenda, entro un'ampia ragnatela ipertestuale, i documenti prodotti da uno specifico soggetto istituzionale e la loro storia, cercando altresì di evitare il più possibile lo stravolgimento dell'identità disciplinare attraverso l'uso di strumenti e classificazioni consueti, in grado di soddisfare le esigenze di rigore interpretativo imposte sia dalla tradizione che dal digitale. Si tratta dunque di un tentativo di *storiografia digitale* che contenga e renda visibili i nessi della documentazione, rispecchiandone il dinamismo e le direzioni strategiche nell'organizzazione politica dell'ente e del territorio, ma i cui connotati principali restino la flessibilità, la pluralità dei percorsi di consultazione e la differenziazione dei possibili approcci.

Jonathan Jarrett

(The Queen's College, Oxford – UK)

Poor tools to think with: the human space in digital diplomatics

The speed of change encapsulated in Moore's Law and its consequences dictate that it is very easy for those who took the time to educate themselves in computing for the humanities to find that it has now moved far beyond them. Projects with cutting-edge methodologies blunt quickly at this pace. Nonetheless, it would seem that in some areas progress is being made more slowly, these being those where the techniques required are more those of cognition than of data management. This paper uses the example of the speaker's own doctoral project, a 'straight' socio-political historical enquiry that happened to employ electronic means of an untheorised kind to manage the diplomatic data on which it was founded, to explore the line between these zones.

When I began my doctorate, which was an attempt to study the workings of power in a frontier area, the *Marca Hispanica* of the Carolingian Empire, my available computing power was a 486 DX-2-66 PC with 8 Mb of RAM and no database software. I was also sharing care of a toddler, while my home institution was in a different city and its computing staff were overstretched. Had my circumstances been otherwise, had I had good technical advice, had I known what I was doing, this paper would probably not exist. I know now that what I wanted was a Wiki-style database with pages for each document, each person, each place, linked to each other by HTML. What I made instead, and still have, was a number of Microsoft Word files peppered with embedded DDE and OLE links to each other. The number of ways in which this is not ideal would be a paper in themselves, but it has one surprising advantage, its subjectivity. Because there is no data schema, I can record what I like in these files, without having to deform the record into shapes determined by a computing structure.

The chief disadvantage was of course that the data capture tools were my eye and my brain, which were not enough. By the time I reached the end of the project I had acquired enough database experience to be working from a reasonably elaborate database in Microsoft Access for certain parts of my sample (some 150 documents) which allowed me to be sure of working with all the data they contained. But this involved huge numbers of decisions as to what each document was about, what role the parties to it were playing, how their names should be spelt and so forth. Since one of the principal purposes of the database was to allow me to decide on the basis of the available information whether the people in the documents, who used no surnames and rarely specified relationships, were recurring, how could I avoid making those choices at data entry stage? This paper uses these questions as a basis for asking larger ones, about the judgements do we ask computers to make, about where human input is still required in digital diplomatics, and whether the advantages of human and computerised judgement can be enjoyed at the same time without diminishment of either?

Luciana Duranti
(British Columbia University – CA)

The Return of Diplomatics As A Forensic Discipline

Research has proven that digital documents, whether born digital or digitized, cannot be preserved. We can only maintain our ability to reproduce them time after time. The most complex aspects of this ongoing preservation involve those activities that aim to counteract system and format obsolescence or to extract documents from their original environment when obsolescence has occurred before any measure could be taken to avoid it. To maintain and assess the authenticity of entities that no longer exist in their native environment requires the strong theoretical and methodological framework which, for traditional documents, has been provided by diplomatics. Although such framework is still valid when examining documents in digital form, it is no longer sufficient, and needs to be integrated with a tested robust practice that allows the certain authentication of what we keep in digital form, such as the digital objects we link to digitised medieval documents to make them accessible and analyse them. This paper will discuss the integration of digital diplomatics with digital forensics, a discipline that originated a decade ago and developed into a rigorous body of concepts, principles and procedures used internationally to fight cybercrime and identify, retrieve, and make accessible authentic digital objects as evidence of the facts and acts they reveal or attest to.

Antonella Ambrosio, Maura Striano
(Università degli studi di Napoli Federico II – IT)

Teaching Diplomatics with New Technologies: a Possible Path

Can we project a Learning Environment for Diplomatics by using new technologies, in order that it could be more effective than in the past traditional way of teaching usual in the lecture rooms? Our proposal aims to answer this question intermixing a typical Pedagogical approach to a Diplomatics one through some experiences we have been carrying out at the University of Naples Federico II. We will focus on how to use a moodle platform,Rete@ccessibile) and the

EditMOM tool online at the website Monasterium.net (http://www.monasterium.net). Rete@ccessibile was successfully experienced in different University courses regarding various subjects and its first goal is to allow full access to every students, paying particular attention to students with (motor, communication, etc.) disability. It aims to be a sort of virtual learning environment but also a place of social interaction and where teachers and students can exchange experiences. EditMOM has a great Diplomatics learning potential which has been experiencing in the last few years during some University courses dedicated to this subject. We analyze the opportunity and the effectiveness to use them for the first time within a Diplomatics course by emphasizing their being complementary.(http://www.firbreteaccessibile.it)

Gunter Vasold

(Karl-Franzens-Universität, Graz – AT)

Online editions as multi-dimensional knowledge spaces

Editorial work leads to a reliable and canonical form of a text. The editor's work is based on the extensive study of the underlying sources and the editor's expertise on the source material. Editing is a dynamic process. But as soon as the edition gets published, it becomes static: the communication between the edition and its user is one-way. Additional insights, new textual traditions, corrections, additional metadata, new objectives and methods or even comments do not have any immediate influence on the edition itself. They may be reflected in other printed publications, or somewhere on the web, but these additions are not available in the context of the original edition. There was some hope that digital editions would provide a better and more dynamic form of interaction. Such editions were called "integrated editions" or "layered editions" whereas I prefer the term *progressive edition*. In such an edition, content (data and knowledge) is iteratively added. This does not only affect the work of the editors, but also leads to better interaction with users, because they can contribute actively via comments or annotations, or implicitly by leaving traces of their usage. A progressive edition is never finished. It represents only the state of knowledge and research interest at a given time. Over time the edition undergoes revisions and additions, made by editors, users, or, as a result of research projects. Imagine specialists in prosopography or toponymy who use data from the edition, work on this data and make their results or even their augmented version of the data known to or accessible from the edition. This could lead to a variety of versions and enrichments from various sources. The resulting edition then becomes a multi-dimensional knowledge space existing of versions, additions and comments of various origins. This data and its relations are hard to manage because of the different versions and also because some comments or additions might only refer to specific versions or variants of the edition. Not only is the management of this data a problem, but so is how to present it to users in a way which makes things transparent but not confusing. Some frequently mentioned problems of digital editions including missing quality control, unclear authority, or long term availability, seem to be worse in such a multi-dimensioned scenario. But I think that there is a way to get things under control, even better than in normal digital editions, if the following conditions are met.

a) It is necessary to put a strong focus on processes. Every step must be (automatically) documented: who did what and when at which stage or version of an edition? In addition every starting and ending point of each process must remain available and quotable. Additionally, transformation rules must be documented and be made automatable as far as possible to guarantee that single processes can be reverted and redone.

b) There must be a central repository containing all stages or layers which have been approved by an editor. This repository can be seen as the authorized part of an edition. It does not only include the authorized data, but should also be able to refer to unapproved additions and comments. These add-ons are not part of the repository, but the repository should be aware of their existence, which would be very helpful for users and editors if they are working on a revised version.

c) To keep things working over a long period, the technical infrastructure must be modular, loosely coupled, distributed and of course based on standards. It has to separate the different areas of origin conceptually, but at the same time it must be able to integrate things in a transparent way from a user's point of view.

d) The repository must be able to deal with complex objects. A diploma edited in this way is more complex than a (structured) text: it has attributes such as images, text versions, all sorts of metadata, references and relations to other attributes and objects and so forth. It should have some sort of methods in the sense of object-orientated programming to act on these attributes, for example to get a specific version of a text or to do an automatic transformation or analysis. Of course it has to be aware of its actual state, and for the sake of flexibility, it should be able to do introspection.

Manfred Thaller

(Universität zu Köln – DE)

What is an Environment for Charters?

In the discussion about tools for IT supported research, the concept of a Virtual Research Environment has recently found considerable attention. It usually describes the concept, that all three major stages of research – the collection of relevant information, its analysis and its publication – should be supported by appropriate computer tools in an integrated way, at the same time be open for collaboration across institutions however.

Many projects of this stripe are intimately bound into national infrastructural programs and therefore extremely general – and therefore so far from actual research, that in the community of computer using Humanists a discussion has recently sprung up, whether this big-project-style of infrastructure-oriented projects is not actually destroying the interest in the Humanities within Humanities Computing.

Using as example a German project for the creation of such a Virtual Research Environment for Charters, the *Virtuelle Deutsche Urkundennetzwerk*, we will try to demonstrate, in which way such an environment can connect closely to concrete research requirements, balancing between the need to create an infrastructure which is general, but at the same time allowing for the simple creation of and interfacing with individual tools, which are very closely modeled on actual research needs.