Journal of Universal Computer Science, vol. 14, no. 2 (2007), 193-210
submitted: 15/6/07, accepted: 17/11/07, appeared: 28/1/08 © J.UCS

Informatics for Historians: Tools for Medieval Document XML Markup, and their Impact on the History-Sciences

Benjamin Burkard
(Hist.-Kulturw. Inf.-Verarbeitung, Univ. zu Köln, Germany
benburkard@gmx.de)

Georg Vogeler
(Histor. Seminar, Ludwig-Max.-Univ. München, Germany
g.vogeler@lmu.de)

Stefan Gruner[1]
(Dept. of Comp.-Science, Univ. of Pretoria, South-Africa
sgruner@cs.up.ac.za)

Abstract: This article is a revised and extended version of [VBG, 07]. We conjecture that the digitalization of historical text documents as a basis of data mining and information retrieval for the purpose of progress in the history sciences is urgently needed. We present a novel, specialist XML tool-suite supporting the working historian in the transcription of original medieval charters into a machine-readable form, and we also address some latest developments which can be found in the field since the publication of [VBG, 07].

Keywords: History-Informatics, Digitalization and Preparation of Medieval Documents for the Semantic Web, XML Tagging, Tool-Support.
Categories: H.1, H.1.m, H.3, H.3.1, H.3.5, H.3.7, H.3.m, H.5, H.5.m, I.7.1, I.7.2, J.5

1 Introduction: Context and History of History-Informatics

History Informatics, previously –especially amongst historians– also known as "Historical Computing"[2] describes an emerging new sub-discipline of Informatics and History, following the examples set by the cross-disciplinary Business Informatics in the 1980s as well as Bio-Medical Informatics in the 1990s. In analogy to those sub-disciplines of Informatics we have recently coined the term "History Informatics", derived from German term "Historische Fach-Informatik" [Tha, 05] which we shall use from this point on. History Informatics belongs to the wider field of "Document-Engineering" and "Digital Humanities"[3] in which much related work can be found, whereby medieval history was the one the first branches of the classical humanities into which computer-based quantitative and statistical research was introduced. The

[1] Corresponding Author
[2] We have abandoned this term, because it could lead to confusion with the study of the history of computer science and historic computing machinery, such as the 1930s and 1940s Zuse automata etc., which is not in the scope of document-processing History-Informatics.
[3] See http://www.digitalhumanities.org/

earliest activities in what we now call History Informatics have been reported already thirty years ago [Bau, 77], which is about a decade after reports about the general possibilities of automated indexing and classification [Bor, 68] as well as automated extracting and abstracting of texts [Wyl, 68] had made their way out of the research laboratories of the System Development Corporation into a more widely accessible monograph. Later, new possibilities of digital representation of source material shifted the focus of History Informatics towards the development of specific database implementations [Tha, 04] and discussions about the possibility of electronic (respectively digital) editions began [SVo, 05]. A comprehensive overview of the field can be found in [BBD, 04].

The difficult relationship between text (syntax) and information (semantics) [BBD, 04] as well as the fact that most interesting information (pragmatics) is not only hidden in but rather "buried under" large amounts of unprocessed text still poses a major obstacle to progress in historical document engineering. For instance, from the European medieval period (500-1500 AD) alone –not to mention other epochs and other regions of the world–, one can soundly estimate several millions of original documents (mostly hand-written on fragile, perishable papyrus or parchment) to be stored in high-security archives, thus largely inaccessible not only to the general public but also to many researchers of medieval history.

Though photographic facsimiles are usually available for the most important bodies of those ancient original document collections, photographic facsimiles are of little use when a historian wishes to full-text-browse, in nowadays fashion, a large archive of medieval documents in search of a particular phrase of text or a particular piece of information, somewhere hidden in a document, anywhere in the entire archive. Therefore, two main aims of History Informatics are:

(i) Provision of methods, techniques and software tools which support the translation of original ancient documents into machine-accessible textual representations (syntax) – our work contributes to *this* field of activity;

(ii) Provision of specialized (topic-specific) information retrieval methods such as indexing, similarity-clustering, data mining or data visualization to make such large amounts of newly generated data automatically accessible to human perception and understanding.

With such digital libraries and semantic retrieval methods becoming more and more available and applicable, we can reasonably expect considerable acceleration in the progress of medieval history research within the next fifteen or twenty years – as we have seen it in Biology and Medicine since they have been supported by Bio- and Medical Informatics.

Whereas photo-optical character recognition [Fel, 01][LMP, 05] in combination with suitable pattern-recognition and machine-learning techniques might possibly be applicable to extract the very *text* of such hand-written medieval documents on a merely lexical level, the *information* inherent to such an ancient piece of text on the level of natural language semantics –which is highly context-sensitive in its various particular historic and linguistic circumstances and thus very much in need of hermeneutic interpretation– can in these days only be extracted by expert historians and philologists specialized in medieval languages. Current work is done in analysis of complex handwritten documents and focuses on image processing. Recent reports

[KKK, 04][HNP, 04] show that already the graphical segmentation of a handwritten text is still a task with many problems to be solved. Basically the same holds true for the project of [BZA, 07] in which Arabic historical manuscripts are being image-processed. At the moment a more promising approach is to support generating data by scholars transcribing and marking up historical texts.

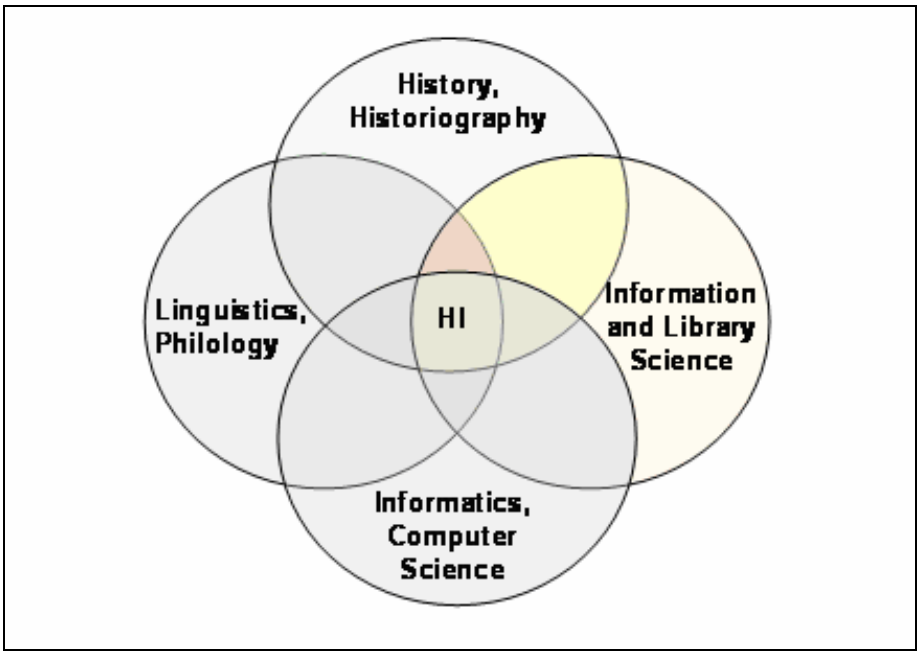

Figure 1: History Informatics (HI) emerging from the intersection of related sciences (History, Historiography, Linguistics, Philology, Library- and Information Science, and Informatics respectively Computer Science)

However, to make such expert-extracted information *explicit* –and thus accessible to the new-generation search engines on the semantic web–, the historian knowledge-engineer must find an adequate way of tagging the digital plain text of those document files with meaningful labels taken from the vocabulary of a both suitable and standardized markup language such as, for example, the widely recognized XML.

This approach was successfully pilot-studied by [Fie, 00] and therefore further promoted in [Vog, 05]. The tedious –yet necessary– transcription task to be performed by the historian knowledge-engineer can, must, and will be supported by specialized markup tools like the ones presented in this paper, which are –to our best knowledge– the *most specialized* markup tools of their kind for the domain of History Informatics (especially as far as medieval European history is concerned). Our tool-suite comprises a purely textual tool as well as a graphically more enhanced tool called EditMOM[4] [Bur, 05] implemented by Benjamin Burkard in the context of the

[4] See http://lehre.hki.uni-koeln.de/EditMOM/ – registration (but no fee) required to enter

Austrian project MOM[5]. Both tools are based on similar principles of tagging which will be further explained in the subsequent sections of this article.

This article extends and revises a presentation which we gave in March 2007 in Seoul, Korea [VBG, 07]. It describes the above-mentioned tools as well as their underlying XML tagging concept, discusses the role of collaborative work in computer based humanities, provides an application example, and discusses why the general (non-specialist) XML tools found in the mainstream of Informatics are of little use for the working historian knowledge-engineer. Such a kind of user is actually interested in ancient documents –not in XML for the sake of its own– and in the impact of the XML tagging concept on historical work; (see [Bel, 06] for a criticism of naïve and inappropriate expectations as far as the general capability of and attitude to XML is concerned). Future work in the History Informatics project is mentioned in the concluding section of this paper. Figure 1 illustrates the relations of History Informatics to other technical and scholarly fields of study.

2 XML Tags for Medieval Charters

A large number of medieval documents belong to the category of *charters*. Those are documents which had particular legal implications (e.g., a contract between a duke and a bishop about the donation of a parish, including the surrounding farmlands, the donation of land at the foundation of a monastery, the grant of the market rights, decisions of a court, a new law, and the like). It is interesting to note that the writers of those ancient charters already had some basic notion of *standardization* such that we can find many charters which are quite similar to each other as far as their *form* and organization is concerned. Due to this technically exploitable formal property [Ans, 99], charters are a suitable starting point as far as document analysis and XML markup is concerned. In a pilot study on charter markup, the *Charters Encoding Initiative* CEI[6], to which Vogeler is affiliated, has started to assemble a preliminary list of XML tags by means of which such charters should be encoded. Thus, that list is not only document-descriptive but also method-prescriptive (until future research results enforce its modification), and our tools are based on those CEI recommendations.

In the remainder of this section we show this list of charter tags and provide some comments and explanations, such that their role in charter document engineering should becomes clear. The tools which support the work with all those tags will be described in the subsequent sections.

2.1 List of CEI Tags, in Alphabetical Order, for Charter Meta-Information

The following XML tags have been proposed by the above-mentioned CEI for the purpose of medieval charter description:

<abstract>, <addressee>, <apprecatio>, <arch>, <archFond>, <arenga>, <auth>, <biblScope>, <CEI>, <chirograph>, <class>, <context>, <corroboratio>, <datatio>,

[5] See http://www.monasterium.net/

[6] See http://www.cei.lmu.de/

<date>, <diplomaticAnalysis>, <dispositio>, <elongata>, <eschatocol>, <formula>, <idno>, <inscriptio>, <insert>, <intitulatio>, <invocatio>, <issuePlace>, <issuer>, <language>, <listBibl>, <material>, <narratio>, <nota>, <notariusSign>, <notariusSub>, <persona>, <petitio>, <pict>, <pictRef>, <protocol>, <publicatio>, <recipient>, <refNum>, <regestum>, <remarks>, <res>, <rubrum>, <sanctio>, <script>, <sigil>, <sigillum>, <subscriptio>, <tenor>, <testes>, <text>, <traditioForm>

Note that some of these tags carry optional attributes such as id=".", lang=".", or type=".", whereas others are atomic and un-attributed. As usual in XML, to every tag of the form <*x*> also the corresponding closing-tag of the form </*x*> exists.

It is easy to see at first glance that this tag language is a *domain-specific* language which could not have been reasonably designed by anybody who is not an expert on medieval charters. This is in contrast to merely editorial or layout markup (such as, e.g., <section> and <subsection> for arbitrary text documents) which can well be automatically recognized and generated by intelligently crafted text processing software like the one we can find in [FGK, 04][Mey, 02]. Standard off-the-shelf XML editors[7] cannot assist the working historian with built-in expert knowledge, and even less adequate for our purpose are other editor tools[8] popular in the field of the clasical humanities that do not allow for any XML tagging at all.

Looking at a given charter, we can distinguish its *form*, its *text*, and its graphical *appearance*. Within the text we further distinguish research-specific contents, and template data. This is reflected in our choice of specialized XML attributes as follows. In a general container of type <text type= "charter"> the charter specific markup highlights formal description characteristics such as: <abstract>, <addressee>, <arch>, <archFond>, <arenga>, <auth>, <class>, <idno>, <traditioForm>, <date>, <diplomatic-Analysis>, <issuePlace>, <issuer>, <language>, <listBibl>, <material>, <recipient>, <refNum>, <regestum>, <remarks>, <res>, <rubrum>, <sigil>, <sigillum>. Text is marked by <tenor>, whereas graphical appearance is described by tags such as <elongata>, <handShift>, <abbr>. Within the text, *specific* content (<persName>, <placeName>, <insert>) as well as the *template* of the text (<apprecatio>, <context>, <corroboratio>, <datatio>, <dispositio>, <eschatocol>, <formula>, <inscriptio>, <intitulatio>, <invocatio>, <narratio>, <nota>, <notariusSign>, <notariusSub>, <petitio>, <pict>, <pictRef>, <protocol>, <publicatio>, <sanctio>, <subscriptio>, <testes>) can also be identified.

2.2 Alternative Tagging Schemas: Related Work

There are other proposals for tagging historical documents, for example the "Encoded Archival Description" EAD[9]. From the more linguistical point of view, the Text Encoding Initiative TEI also deserves to be mentioned. Both these approaches have their advantages, as discussed in the following.

EAD follows a hierarchal concept of archival management. It uses the hierarchical concept of XML to build containers for archival fonds and abstract items

[7] See for example http://www.stylusstudio.com/

[8] See for example http://karas.ch/cet/ or http://www.oeaw.ac.at/kvk/cte/main.htm

[9] See http://www.loc.gov/ead/

that can contain other archival items in a recursive way via a component element <*C*>. The description of individual items in EAD schemas is more cursory. Definitively beyond the usage scenario of the EAD is the description of the actual text. For the historical research, EAD is only a schema delivering metadata – not the object data itself. Exchange with archival data is a crucial point for the success of any tool using historical material, as the archives keep the original documents and provide their meta data. The CEI Working Group thus tries to keep in contact with the archival scene. The main strategy to keep both concepts close to each other is to propose the CEI element names as a kind of a controlled vocabulary that can be used to specify the generic concept of the <*C*> component when applied to medieval legal documents.

TEI[10] offers a large set of XML tags for the encoding of linguistic phenomena. Its latest version[11] P5 comprises tags for manuscript description, too. Comparing this tag set with the tag set of CEI we can observe a considerable overlap of both tagging schemas. The CEI schema proposes some more specific elements that can be understood as instances of more generic elements. For example, <testis> is an instance of <persName>. Moreover, the elements describing the Diplomatic discourse (<intitulatio>, <arenga> etc.) could be labelled by generic segment elements (<seg>), specified by type or function attributes. However CEI does not stop at defining specific instances of linguistic concepts which could also be expressed in the language of TEI. In the CEI approach it is possible to add concepts of *transmission* and *authentication* which are of general importance in document management and thus specifically for historical documents, too. As the TEI community has always been open-minded towards new concepts during the past twenty years, it seems reasonable to conjecture that the schemas of CEI and TEI may once be merged. In fact the integration process has already started with the development of CEI's new ODD[12].

3 Tool Support

This section describes the software tools which we have currently available (in stable versions) for the purpose of XML tagging of medieval charters, and illustrates their application by example. Recently emerging issues and future developments are addressed in the subsequent sections.

3.1 Textual Prototype

The simplest tool in our tool-suite is a mixed-mode text editor, the interface of which is shown in figure 2. It supports the preparation, creation and modification of XML-tagged text files as "semantic copies" of medieval charters for the purpose of storing them in databases and accessing them via the new-generation semantic web [SVo, 05][Uhd, 99]. Whereas the lexical contents of the original charter must be manually typed in free text mode –yet somewhat similar to the work of a medieval copyist

[10] See http://www.tei-c.org
[11] Release announced for the last quarter of 2007
[12] To appear on the CEI website – see footnote 6

sitting in his modestly equipped scriptorium– and cannot yet be imported by OCR scanning, the selection of the XML tags (to markup the to-be-digitalized charter's text) is menu-driven. Note that the tag menu is *context-dependent*, which means that not every tag is available for use in every arbitrary situation. For example, while an <abstract> tag is open, only the tags <issuer>, <geogName>, <persName>, <recipient>, <addressee> and/or <placeName> may be selected, and all other tags cannot be chosen in that situation.

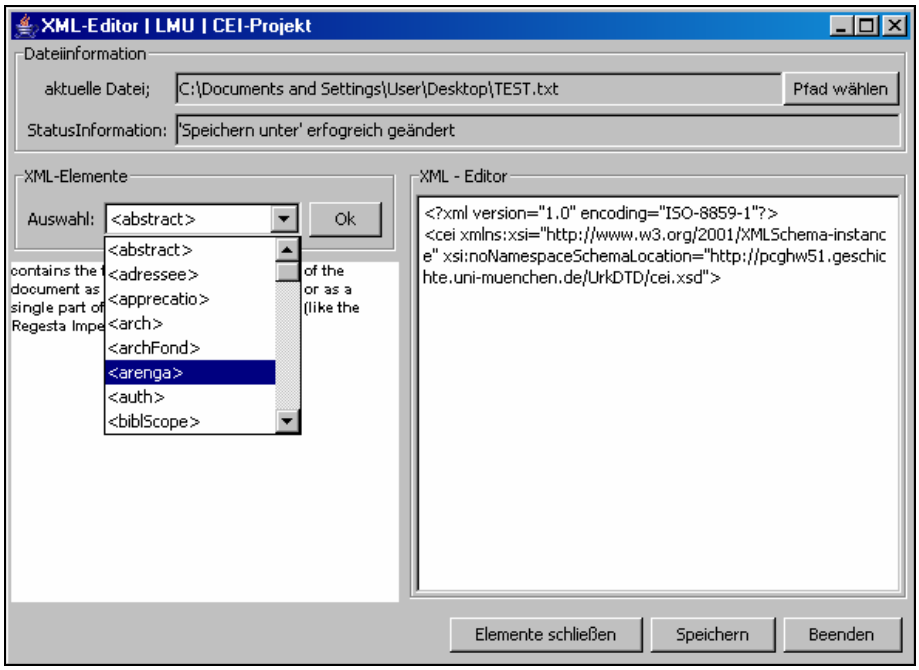

Figure 2: Textual Tagging Tool, with Context-Dependent Tagging Menu

Technically, the underlying XML grammar defines those contextual relationships. This feature is supposed to minimize the possibility of wrong tagging and is justified by the above-mentioned standardized structure of the charters we are dealing with. Thus, not only the topic-specific XML tags as such but also their meaningful relationships amongst each other are implemented into the tool as an *implicit ontology*.

3.2 EditMOM

A more advanced implementation of these concepts was added to our tool-suite in form of EditMOM in the context of the already mentioned Monasterium project MOM [Bur, 05]. The two major features of this tool are the following:

(i) Textual tagging, as shown in figure 2 is replaced by *graphical symbols* which are more intuitive and easier to use by historians and literature

scholars unfamiliar with the technicalities XML. Markup symbols can be reduced to small arrows, or can be expanded with a caption, yielding a common term which is known to the working historian.

(ii) The elements are internally mapped to specific parts of the GUI, such as: continuous text editor, attribute masks, or table-formed lists.

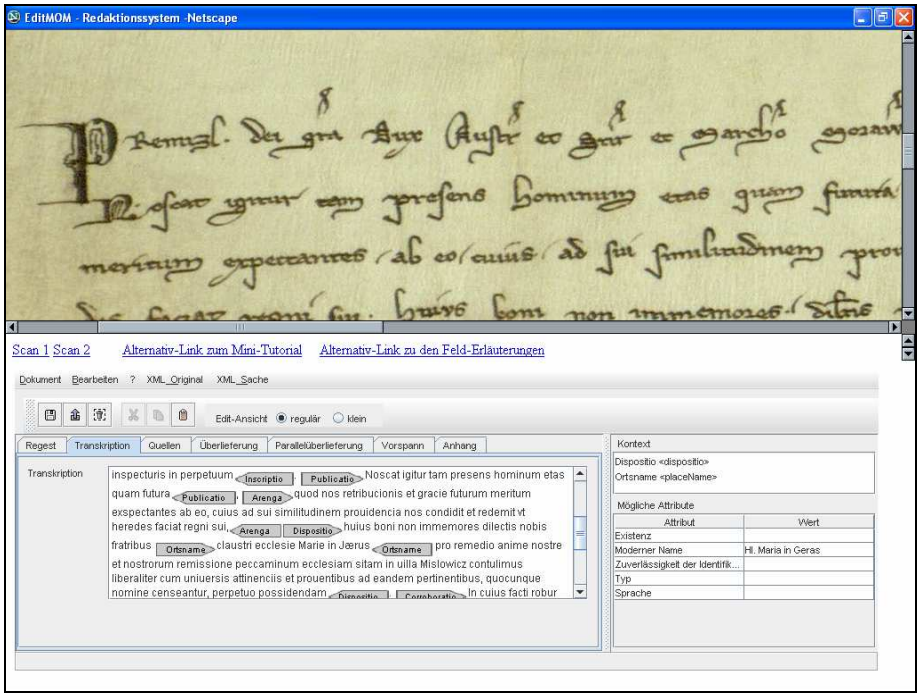

Figure 3: EditMOM, with Graphical Features and Photographic Charter Facsimile

Figure 3 shows the EditMOM tool in combination with a photographic scan of an original document which the historian is supposed to read and see while creating its XML representation through the GUI of this tool. Thus, the working historian can use the tool through different *views* at different levels of abstraction –some of which hiding the details, others exposing them–, anyway providing fields for attributes, context information, and the like.

Technically, a specific definition meta-file "EditMOM.xml" relates the various view properties of the markup to each other and groups them into menus, tabsulators (etc.), as shown in the (rather textual) figure 4. We can also see in figure 4 that the definition file contains translated terms, thus the tool finally serves as a language- and domain *specific* editor for the working historian who is dealing with medieval (including early-modern) charters. Note that it is especially the <doc> sections of the EditMOM definition file that provide advice and instructions for the working historian.

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Version 0.9 2006-08-10 , by: GV -->
<EditMOM xsi:noNamespaceSchemaLocation="EditMOM.xsd" xmlns:xsi=
"http://www.w3.org/2001/XMLSchema-instance">
<tab>
<name>Regest</name>
<caption lang="de">Regest</caption>
<caption lang="en">Charter-Description</caption>
<field>
        <name>idno</name>
        <path>/text/body/idno</path>
        <present_type>mask-field</present_type>
        <rowCount>1</rowCount>
        <caption lang="de">Signatur</caption>
        <caption lang="en">ref.- num.</caption>
        <doc lang="en">must contain the ID number of an object within its current context
(e.g. book of charters), its location and label in an archive, prominent characteristics which
distinguish this object from other objects, date, age or period, etc. </doc>
</field>
<field>
        <name>Abstract</name>
        <path>/text/body/chDesc/abstract</path>
        <present_type>Free-Text</present_type>
        <rowCount>5</rowCount>
        <caption lang="de">Regest</caption>
        <caption lang="en">Abstract</caption>
        <doc lang="en">this must contain the Regest (abstract) which summarizes the legal
contents of a charter document. May be a short header-Regest in editions, may also be a
detailed long-Regest in other contexts (including verbatim quotations). Note: this abstract
MUST mention the author/sender of the charter and SHOULD mention its receiver.</doc>
<attributes>
        <attribute>
        <name>lang</name>
        <caption lang="de">Sprache</caption>
        <caption lang="en">language</caption>
        <doc lang="en">contains an abbreviation (such as: de, la, mhd, hu), in case that the
object languages deviates from the usual language of the context in which the object is
embedded.</doc>
        </attribute>
</attributes>
</field>
```

Figure 4: EditMOM Meta File, defining Charter Representation and related Features

All in all it is the objective of the EditMOM tool to meet the specific requirements of the practically working historian. It seems important to support him (or her) in a manner that connects to the way he used to do his work in pre-electronic time: Just like in his traditional work on medieval charters the historian can view the image of the charter while collecting data about it. Also, in order to get to know

content and structure of charters the highlighting of certain passages (e.g. index terms such as persons or specific structures) has been an integral part in traditional education of historians specializing in medieval history. This exact method is used with EditMOM when the user highlights a passage inside the text and then selects the adequate element. At the same time a strong emphasis has been put on keeping the editor as easy to use as possible, as most historians even nowadays still have only limited experience with computers. Thus, among others, the XML tags are hidden from the user, context-sensitive tagging is provided and overlapping of XML structures are prevented.

3.3 Application Example

In this sub-section we illustrate step by step an application of EditMOM. Starting with a photographic facsimile of a genuine, material medieval charter we obtain an XML-tagged ASCII-version of it such that its non-material essence is extracted from its perishable matter for a future existence in the virtual world of the semantic web. Once the charter's XML file is created, its tags may not only be used for extracting semantic information but also for rendering a new image of that charter on the screen, similar to the graphical interpretation of traditional HTML documents.

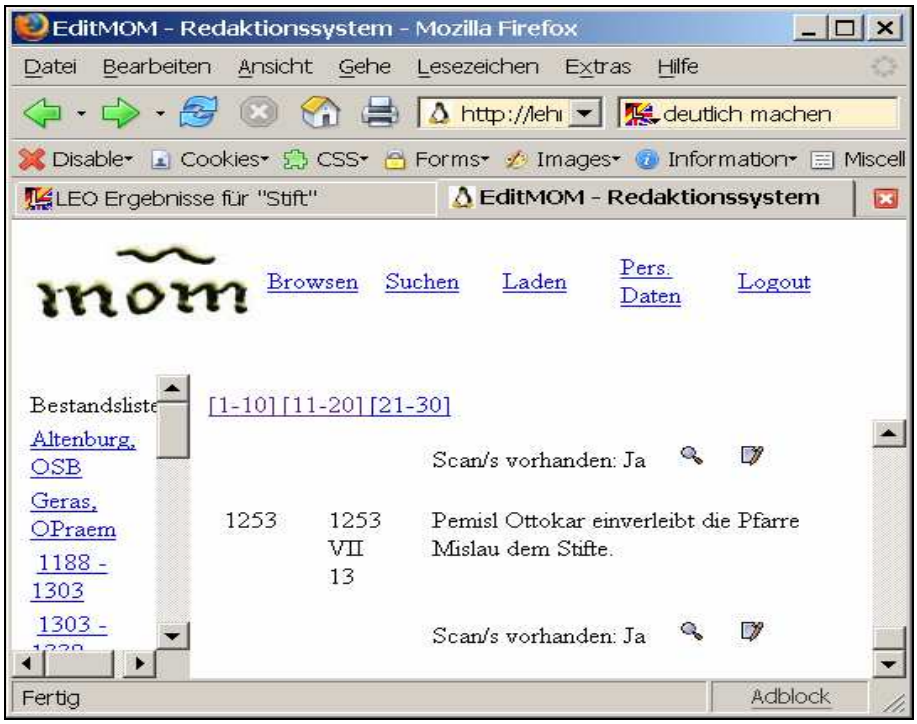

Figure 5: Integrated Web-Application, providing access to a pool of photographic facsimiles of Charters which are yet to be edited with help of the XML Tagging Tool

Assume that a working historian wishes to document-engineer a charter which has not yet been transformed to XML-tagged plain-text. The MOM web application shown in figure 5, also developed by Burkard, provides a keyword-searchable database of such charters together with the editing tool of figure 3. Thus, via this web application the working historian can search and access the photographic images of charters and use the integrated XML tagging editor EditMOM at the same time. This integrated approach improves the working conditions for the individual historian as well as it supports cooperation amongst a group of historians via the Internet.

Now assume that the historian has chosen the charter shown in figure 6 for translation into XML-tagged plain-text. It is a facsimile of a charter by Duke *Premisl Ottokar* who stipulated, by means of this charter, that the Parish of Mislau be transferred to the dominion of the Monastery of Geras in Lower-Austria in the year 1253. A brief summary of the contents of this charter can be seen in the interface of figure 5 where it says in German tranlation of the Latin original: *Premisl Ottokar einverleibt die Pfarre Mislau dem Stifte*, whereby the word *Stift* (= dominion of a monastry) had been entered by the user as a keyword into the search-facility of the web-interface: see *LEO Ergebnisse* (= results) *für Stift* in figure 5. The material original of this charter can be found in the archives of that monastery where also the photographic image was taken. This charter shown in figure 6 is part of the already mentioned Monasterium project [Aig, 03] in the context of which the EditMOM tool has already demonstrated its usefulness.

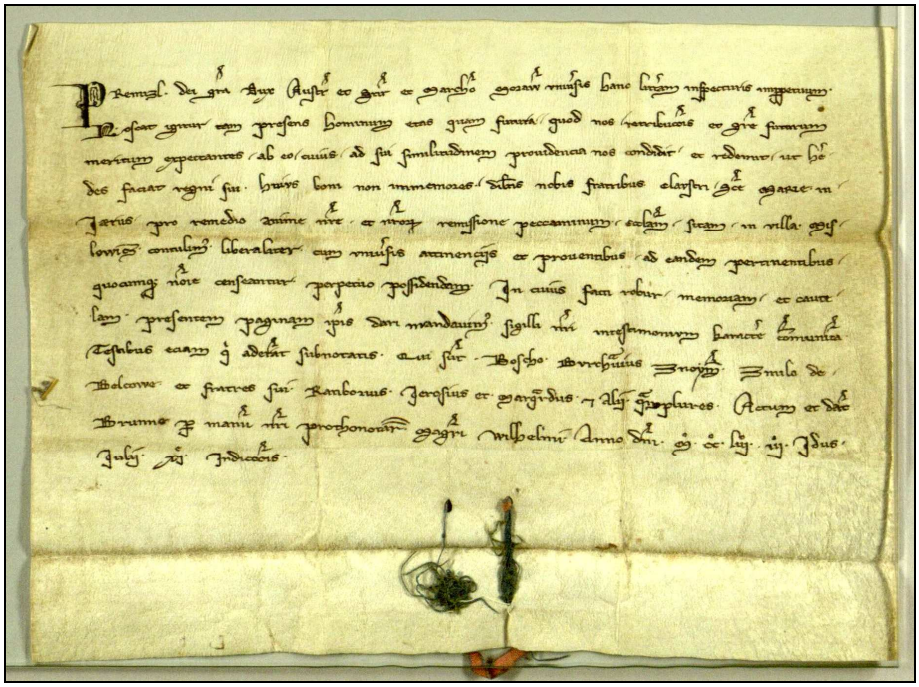

Figure 6: Premisl Ottokar's Charter about Mislau [Source: Monasterium.NET]

```
<?xml version="1.0" encoding="utf-8">
<cei><text type="charter">
<tenor>
<intitulatio>
        <persName reg="Otokar (II.), Herzog von Österreich und Steiermark (1251-1276), Aussteller">Premizl, dei <expan abbr="gra">gracia</expan> dux <expan abbr="Aust">Austrie</expan>et<expan abbr="Stir">Stir</expan>ie et <expan abbr = "marcho">marchio</expan> <expan type = "Moraw"> Morawie </expan> </persName>
</intitulatio>
<inscriptio>
        vniu <expan abbr = "er-Haken">er</expan> sis hanc litt <expan abbr = "er-Haken"> er </expan>am inspecturis in perpetuum
</inscriptio>
<publicatio>Noscat igitur tam presens hominum etas quam futura</publicatio>
<arenga>
        quod nos retribucionis et gracie futurum meritum exspectantes ab eo, cuius ad sui similitudinem prouidencia nos condidit et redemit vt heredes faciat regni sui,
</arenga>
<dispositio>
        huius boni non immemores dilectis nobis fratribus <placeName reg="Hl. Maria in Geras">claustri ecclesie Marie in Jærus</placeName> pro remedio anime nostre et nostrorum remissione peccaminum ecclesiam sitam in uilla Mislowicz contulimus liberaliter cum uniuersis attinenciis et prouentibus ad eandem pertinentibus, quocunque nomine censeantur, perpetuo possidendam
</dispositio>
<corroboratio>
        In cuius facti robur memoriam et cautelam presentem paginam ipsis dari mandauimus sigilli nostri in testimonium Karactere communitam. Testibus eciam qui aderant subnotatis; qui sunt
        <testis>Bosko Burgrauius Znoymensis</testis>,
        <testis>Zmilo de Belcowe et fratres sui Ratiborius</testis>,
        <testis>Jerosius</testis> et
        <testis> Marqvardus </testis> et alii quam plures
</corroboratio>
<datatio>
        Actum et datum Brune per manum nostri prothonotarii magistri Wilhelmi. Anno domini <num type = "römische Zahl" value = "1253">M. CC. LIII.</num> <num type = "römische Zahl" value = "3">III</num> idus Julij <num type = "römische Zahl" value = "11">XI </num> indictionis.
</datatio>
</tenor>
</text>
</cei>
```

Figure 7: XML-tagged plain-text Transcription of Ottokar's Charter

In the tool's editor, the working historian would then add the formal description requested by the above-mentioned EditMOM definition file (i.e.: abstract, remarks on the language of the charters, etc., as shown in figure 4) and then transcribe the original text from the photographic image into the textual input window of the XML editor. The tool supports this tedious scholarly work with special options and menus for the markup (tagging). These tagging-aids and editing-features carry intuitive names from the expert's domain terminology of historical research with which the historian is familiar. The result of the tool-assisted transcription of Ottokar's charter is the well-formed and parse-able XML file shown in figure 7. Moreover: this kind of markup also enables the scholar to search for the use of abbreviations, or to search for a person with a later given modern name that differs completely from his original historical name. It is now also possible to compare typical *patterns*, or parts thereof (as denoted by the <arenga> sections), where the medieval scriptorium-clerks professionally reverted to quasi-standardised examples of good practice which they re-used and customized to the purpose of their subsequent charters in-the-making[13]. Such kind of queries can obviously be machine-supported, such that the historian would have access to information on the activities of *Ottokar* without the need to know any more about computer technology than web-surfing as well as the common use of interface-windows and user-menus.

4 Recent Issues

A first usability test undertaken in July 2006 has indicated that a group of twenty historians, with only limited computer skills but specialists in the field of medieval charters, were able to work with the tool without any major complications, and further usability-related investigations of this kind are planned for the future. Since that usability test, the EditMOM tool has grown into a productive stage since December 2006. The data migration of more than 100.000 charters in the relational database of the Monasterium project is on its way and is scheduled to be finished by the end of 2007. Currently there are about hundred registered users who use the tool in a mostly academic environment. Moreover, the tool is Open-Source under the terms of GNU General Public License since the early months of 2007. The integration of the XML-tagging support (by means of source code reuse) into other projects, as well as contributions to integrating external data resources are desired because of various methodical and semantic issues in those related fields which seem to be suitable for tool support, too.

4.1 Collaborative Work

The growing number of EditMOM's registered users leads to the question how quality can be controlled in a humanities computing project like the Monasterium project. One goal of the EditMOM tool is to fructify the large public interest in local

[13] Needless to say that Duke Ottokar did not write all his charters with his own hands: that was the job of the writers working in his chancery. Document experts in this field can recognize and distinguish the regional, formal and stylistic subtleties related to the various episcopal or principal chanceries operating in those times.

history, making our cultural heritage accessible via the internet through a large number of geographically separated contributors. Consequently, a more collaborative approach of work should be supported, similar to what we have already seen in the now well-established field of tool-supported Collaborative Software Engineering [Gol, 02] as pioneered since about twenty years ago by Finkelstein et al. As far as EditMOM is concerned, this is ongoing work at the University of Cologne. For the purpose of quality assurance, EditMOM-Tool follows a workflow model defined in [Bur, 07] which invests the authority of final publishing decisions onto a user group called "moderators". Ordinary users have to register before being allowed to edit a charter description, and only after moderator approval a digitalized charter can be released to the public domain.

The main purpose of the moderator view is to provide a simple and easy-to-use overview of the changes (differences) made by ordinary users during their daily or sporadic activities. As the EditMOM tool hides XML-Code and breaks complex structures into various fields, the presentation of the differences are currently structured by the already mentioned tabulators. For editor fields containing hidden XML-tags, however, the comparison view must be aware of the underlying XML. Currently existing XML-diff-tools such as [HMe, 05] do not sufficiently meet the requirements of this scenario, as they are merely syntax-driven and thus not able to exploit subject-specific (semantic) information from the given application domain. Ongoing work in the EditMOM project is aimed at such "semantic" tools.

4.2 Methodological and Inter-Regional Effects and Implications

Science is and has always been an international endeavour – only its lingua franca has been replaced by another one every now and then. This has implications on a web-based tool suite like MOM/EditMOM which could, in principle, be accessed and used from any place in the world, where a minimum of digital communication infrastructure is available. For this reason there are ongoing activities of translating the currently only German-language user-interface into other Central European languages, such as Czech, Slovenian, Slovakian, Hungarian and Italian. Immediately the question arises: why these "small" languages – why not (yet) Spanish, Arabian, Chinese or the current lingua franca: English? The answer to this question is two-fold:

(i) The above-mentioned Austrian Monasterium project, for which our tools had been primarily been developed, is dealing with charters issued by historical figures (like the above-mentioned Duke Ottokar) who's political activities reached out across a Central / Eastern / Southern European region which overlaps –in total or in parts– the territories of the above-mentioned modern countries. From this regional research point of view is self-understanding that EditMOM's interface language support for Czech or Hungarian is of higher priority than for English or Spanish – for countries which figures like Duke Ottokar did not reach.

(ii) For similar region-historic reasons, the English language as the current lingua franca of science has not yet fully developed a highly specialized terminology which is needed to adequately describe and communicate

the topic-specific details and subtleties of Central European medieval charter research – an old field of study which had already developed a sophisticated terminology long before English became the global lingua franca of science, especially after WWII.

It is thus interesting to note that software tools like EditMOM do not only support the working historians in their daily tasks, but –by enabling inter-regional collaboration through the internet– also bring to light the discrepancies of regional scholarly terminologies. The dissemination and increased availability of collaborative research support tools like EditMOM can thus be regarded as a driving force behind an accelerated scholarly discourse towards the semantic clarification and assimilation of disparate regional terminologies. A process that has not yet been fully understood can also not be fully automated, which means (vice versa) that difficulties in the provision of tool support –even if it is only the difficulty of translating the widgets of a user interface from one natural language into another one– might be regarded as evidence of a certain gap in the factual or methodological understanding of the matters at hand.

4.3 Future Work

As far as future work is concerned we shall mention improvements of our prototypes themselves as well as novel applications of them. As mentioned above, the manual copy-typing of an original charter's contents in a tool's free-text editor is –apart from the menu-supported XML tagging– rather tedious and "medieval" task, such that a future tool should be coupled with an optical scanner to read in the text from a photographic facsimile [Fel, 01] of the charter to be transcribed. As far as the <language> attribute is concerned, it should be possible to automatically recognize a charter's source language, e.g. Medieval Latin or Middle High German, by means of the same methods that are usually applied in automated recognition of modern languages: see for example [GLi, 04] [MSi, 05]. Moreover, the XML tag set could be augmented not only to capture further features of charters, including material ones like <color> or <parchment>, but also further classes of medieval documents other than charters, for example medieval poetry, medieval science-books, and the like.

5 Summary and Conclusions

The way of delivering original documents to historians are scholarly editions. The long tradition of producing such editions has developed useful techniques and structures for transcribing and describing, distinguishing forgeries, summing up main issues, dating and locating, marking linguistic and content structures in the text, etc. These well established techniques are yet not very well supported by the computer, as common text processing software only aims at standard office applications. The recent developments of markup techniques promise major improvements but have not been very much used by the historical scholars as there are no suitable user interfaces supporting the well-established methods of historical work. We thus want to build tools that can be used intuitively by scholars familiar with no more than everyday office software. The EditMOM tool should be used by any scholar trained in working

with original texts from ancient periods, providing a working environment that is based on the concepts of his discipline. The prototype was developed for medieval charters, but its GUI concepts can, in principle, fit into the work with other kinds of sources (e.g. letter collections, files, antique inscriptions, pragmatic manuscripts, financial records etc.) if appropriate underlying XML schemas are made available to support the purpose. Based on the assumption that the digitalization of historical text documents as a basis of data mining and information retrieval for the purpose of progress in the history sciences is urgently needed, we have presented a novel, specialist XML tool-suite the tools of which support the working historian in the transcription of original medieval charters into a machine-readable form. Future applications of such tools are likely to lead to an accelerated growth in the number of marked-up historical documents available in specialist library databases, awaiting further automated knowledge extraction by all available means [GMa, 03][HVo, 05] of information retrieval, logics/statistics-based automated reasoning or data mining. In consequence, accelerated progress in the history sciences can reasonably be foreseen.

Acknowledgements

Thanks to Martin Gruner for the implementation of the textual tool. Thanks to Manfred Thaller for establishing this project as well as for fruitful discussions since then. Thanks to Mirko Gontek for his ongoing contributions to the further development of the EditMOM tool. Thanks to Rafael Lins for organizing an interesting event as well as for inviting our contribution to this special issue of JUCS. Last but not least thanks to Hermann Maurer, editor of JUCS, for publishing our article with its obvious links to the history of his native land.

References

[Aig, 03] T. Aigner, Strategien zur digitalen Bereitstellung historischer Quellen aus den Archiven der nieder-österreichischen Ordensstifte. Archive und Forschung, Der Archivar, Beiband 8, pp.295-306, Siegburg 2003.

[Ans, 99] M. Ansani, Diplomatica e diplomatisti nell'arena Digitale. Scrineum 1, pp.1-11, 1999.

[Bau, 77] R.-H. Bautier, Les Demandes des Historiens à l'Informatique La Forme Diplomatique et le Contenu Juridique des Actes. Informatique et Histoire Médiévale: Actes du Colloque de Rome, 20-22 May 1975. Publications de l'Ecole Française de Rome 31, pp.179-186, 1977.

[BBD, 04] O. Boonstra, L. Breure, P. Doorn, Past, Present and Future of Historical Information Science. Technical Report: Koninkl.-Nederl. Akademie van Wetenschappen, Amsterdam 2004.

[Bel, 06] A. Bell, Software Development amidst the Whiz of Silver Bullets. ACM Queue 4/5, June 2006.

[Bor, 68] H. Borko, Indexing and Classification. In H. Borko (ed.), Automated Language Processing, pp.99-125, John Wiley and Sons Inc. publ., New York, October 1968.

[Bur, 05] B. Burkard, Collaboration on medieval charters – Wikipedia in the Humanities? Proc. 16th Conf. of the Association for History and Computing, pp.91-94, Amsterdam, 2005.

[Bur, 07] B. Burkard, Wiki goes Humanities. Kollaborative Erschließung mittelalterlicher Urkunden, in: Wikis im Social Web - Wikiposium 2005/06, pp.130-144, Wien/Vienna 2007.

[BZA, 07] W. Boussellaa, A. Zahour, A. Alimi, A Methodology for the Separation of Foreground / Background in Arabic Historical Manuscripts using Hybrid Methods. Document

Engineering Track: SAC'07 Annual ACM Symposium on Applied Computing, Vol.1, pp.605-609, Seoul, Korea. ACM Press, March 2007.

[Fel, 01] B. Feldmann, OCR von Handschriften. Codices Electronici Ecclesiae Coloniensis: Eine mittelalterliche Kathedralbibliothek in digitaler Form. Fundus Forum für Geschichte und ihre Quellen, Beiheft 1, pp.107-143, Göttingen 2001.

[FGK, 04] C. Fuß, F. Gatzemeier, M. Kirchhof, O. Meyer, Inferring Structure Information from Typography. Digital Documents: Systems and Principles, pp.44-45, LNCS 2023, Springer-Verlag, 2004.

[Fie, 00] A. Fiebig, Urkundentext: Computergestützte Auswertung deutschsprachiger Urkunden der Kuenringer auf Basis der eXtensible Markup Language (XML). Schriften zur süd-westdeutschen Landeskunde 33, Leinfelden-Echterdingen 2000.

[GLi, 04] P. Goncalves, R. Lins, Automatic Language Identification of written Texts. Proc. SAC 2004 Annual ACM Symp. on Appl. Comp., Document Engineering Track, Nicosia, 2004.

[GMa, 03] M. Gervers, M. Margolin, Application of Computerized Analyses in Dating Procedures for Medieval Charters, Le Médiéviste et l'Ordinateur 42, 2003.

[Gol, 02] A. Goldberg, Collaborative Software Engineering. Journal of Object Technology Vol.1, No.1, pp.1-19, May/June 2002.

[HMe, 05] D. Hottinger, F. Meyer, XML-Diff-Algorithmen, Techn. Report, ETH Zürich, July 2005, http://www.infsec.ethz.ch/education/projects/archive/XMLDiffReport.pdf

[HNP, 04] L. Heutte, S. Nicolas, T. Paquet, Enriching Historical Manuscripts: The Bovary Project. Proc. DAS'04, pp.135-146, LNCS 3163, Springer-Verlag, 2004.

[HVo, 05] M. Heller, G. Vogeler, Modern Information Retrieval Technology for Historical Documents. Proc. 16th Conf. of the Association for History and Computing, pp.143-148, Amsterdam, 2005.

[KKK, 04] H.J. Kim, M.S. Kim, H.K. Kwag, K.T. Cho, Segmentation of Hand-Written Characters for Digitalizing Korean Historical Documents. Proc. DAS'04, pp.114-124, LNCS 3163, Springer-Verlag, 2004.

[LMP, 05] G. Leedham, K. Melikhov, V. Pervouchine, Handwritten Character Skeletonisation for Forensic Document Analysis. Proc SAC 2005 ACM Symp. on Appl. Comp., Track Document Engineering, Santa Fe, 2005.

[Mey, 02] O. Meyer, aTool: Creating Validated XML Documents on the fly using MS Word. Proc 20th internat. Conf. on System Documentation, Toronto, pp.113-121, ACM Press, 2002.

[MSi, 05] B. Martins, M. Silva, Language Identification in Web Pages. Proc SAC 2005 ACM Symp. on Appl. Comp., Track Document Engineering, Santa Fe, 2005.

[SVo, 05] P. Sahle, G. Vogeler, Urkundenforschung und Urkundenedition im Digitalen Zeitalter. HIST2003 Geschichte und Neue Medien in Forschung, Archiven, Bibliotheken und Museen. Historisches Forum: Schriftenreihe Clio-Online 7/1, pp.333-382, Berlin 2005.

[Tha, 04] M. Thaller, Texts, Databases: a Note on the Architecture of Computer Systems for the Humanities. Augmenting Comprehension: Digital Tools and the History of Ideas. Office for Humanities Communications Publications 17, pp.49-76, London 2004.

[Tha, 05] M. Thaller, Historische Fachinformatik: ein Kölner Modell. HIST2003 Geschichte und Neue Medien in Forschung, Archiven, Bibliotheken und Museen. Historisches Forum: Schriftenreihe Clio-Online 7/1, pp.45-72, Berlin 2005.

[Uhd, 99] K. Uhde, Urkunden im Internet: Neue Präsentationsformen alter Archivalien. Archiv für Diplomatik 45, pp.441-464, 1999.

[Vog, 05] G. Vogeler, Towards a Standard Encoding of Medieval Charters with XML. Literary and Linguistic Computing 20, pp.269-280, 2005.

[VBG, 07] G. Vogeler, B. Burkard, S. Gruner, New Specialist Tools for Medieval Document XML Markup. Document Engineering Track: SAC'07 Annual ACM Symposium on Applied Computing, Vol.1, pp.594-599, Seoul, Korea. ACM Press, March 2007.

[Wyl, 68] R.E. Wyllys, Extracting and Abstracting by Computer. In H. Borko (ed.), Automated Language Processing, pp.127-179, John Wiley and Sons Inc. publ., New York, October 1968.