XML TO VISUAL TAGS MIGRATION
PROPOSED METHODOLOGIES FOR THE RESEARCH OF DIGITISED CROATIAN MEDIAEVAL GLAGOLITIC MATERIAL

Mario Essert
Faculty of Mechanical Engineering and Naval Architecture, Department of Control Engineering, University of Zagreb, Croatia

Boris Bosančić
Faculty of Philosophy, Department of Library and Information Science, Josip Juraj Strossmayer University of Osijek, Croatia

Marijana Tomić
Department of Information Sciences, University of Zadar, Croatia

Mario Lončarić
Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia

Abstract

Manuscripts are often fundamental sources for researchers in many disciplines, and nowadays this research is enhanced by new tools brought up by information technology. In this article, different methodologies of manuscript research in digital environment based on annotation are described. In the first part of the article, the annotation itself and the XML (eXtensible Markup Language) are explained. In the second part, different types of annotation methodologies are proposed for two types of researches, TEI and TEIMark that serve for encoding of transcribed versions of manuscripts and are proposed for researches based on a text, and DocMark that serves for encoding and analysing of digitised versions of manuscripts and is proposed for researches based on a text together with its physical context. Additionally, History Integrator software which places manuscripts into the time and space framework, with additional multimedia information is described.

Keywords

digital humanities, annotation, XML, TEI, image markup tool, web-based encoded documents analysis, Croatian Mediaeval Glagolitic material

Introduction

Manuscripts are often fundamental sources for researchers in many disciplines, for example in palaeography, codicology, linguistics, history, art history, musicology etc. Therefore, manuscript heritage collections are very important and need to be preserved, protected, described and made available for research. Those tasks are nowadays enhanced by information technologies which provide new possibilities of this research. It is obvious that the demands of the humanities nowadays stimulate the transformation of the technologies, and vice versa, the technologies change and direct humanities' research. This is the case with manuscript studies also, as they are enriched by new tools and new possibilities that provoke and direct new researches and methodologies.

One of the best known and used methodologies for manuscript research in digital environment is TEI annotation, based on XML, which we will present among other methodologies in this article. The prerequisite of using TEI is to transcribe a document. Once a transcript is available, the user can encode a text. Mostly, it is used for linguistic, text based research, as indicated by the name (*Text Encoding Initiative*). Despite many advantages of that methodology, the loss of context of a researched text in terms of visibility of characteristics of codices as physical objects can disable certain types of research based on manuscript as material, and not just linguistic objects. The examples are some types of palaeographical or codicological researches that need an insight into a material context of a text like the research of *mise en page*, morphology of letters, distinguishing among various hands on the same manuscript, etc. That approach is described by Elena Pierazzo and Peter A. Stokes who uphold a codicological approach to manuscript research led by the idea of "putting the text back into context" and stop forcing the scholars "to consider the text first".[1] Pierazzo and Stokes brought to mind that "the fact that the text was transmitted to us by means of a specific physical object which has been organised in a certain way and preserved in one place or another has all sorts of consequences in the way we understand and receive that text."[2] They uphold the idea that for the understanding of a text in a manuscript, it is fundamental to

1 See more on the subject in: Stokes, Peter A.; Elena Pierazzo. Putting the text back into context : a codicological approach to manuscript transcription. // Codicology and paleography in the digital age 2 / edited by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt : Herstellung und Verlag, 2010. Pp. 397-430.

2 Ibid., p. 398.

study the layout, the type of script, the type of writing support, the binding and other physical elements that are usually main features for identifying the scribe, scriptorium, date and place of production, etc. Following that idea, we propose a methodology of manuscript research based on annotation of digitised images, instead of transcribed texts, which is appropriate for the research of features dependent on physical, rather than linguistic characteristics of manuscripts. The tool that enables this methodology will also be presented in this article. Finally, History Integrator, a computer program that visualises information on researched manuscript by placing it in the context of time and place of its production, recording the history of its movements and present-day location, will be described.

Annotations tools and History Integrator visualisation computer program are examined in this article in respect of research of specific Croatian mediaeval corpus, as well as in respect of general features concerning manuscript research. In addition to usual, expected issues occurring during manuscripts research in the digital age, like low readability etc., a researcher of Croatian mediaeval corpus encounters a range of specific problems resulting from the frequently quoted difficulty due to the use of three languages and three alphabets. The corpus was written in three alphabets: Glagolitic, Latin and Bosnian variant of Cyrillic scripts, called *bosanica* or *bosančica*, and in three languages: Church Slavonic, Croatian redaction of the Church Slavonic language and Latin.[3] Since the Croatian Mediaeval Glagolitic Corpus is in the focus of this article, we should mention the evolution of Glagolitic script from its rounded form to the specific Croatian type, angular Glagolitic script, as well as various inter-scriptural and other cultural influences which had diverse effects in different text types and thus affected the modifications in the morphology of letters, page setup, the use of certain linguistic elements, book format and alike.[4] The above mentioned diversity of Croatian Mediaeval scripts and letter forms in each script heretofore prevented former attempts of building OCR (Optical Character Recognition) programs that could be applied to Croatian Glagolitic manuscript material, so the transcription of that material can be hard and time consuming.

3 More on the subject see in: Hercigonja, Eduard. Trojezična i tropismena kultura hrvatskog srednjovjekovlja. 2. dopunjeno i izmijenjeno izd. Zagreb : Matica hrvatska, 2006.

4 Žagar, Mateo. Osnovni procesi konstituiranja ustavne glagoljice. // Вългари і Нърват prez vekovete II / ur. Rumjana Božilova. Sofija : IK Gutenberg, 2003. P. 31.

Some researchers of Croatian Glagolitic manuscripts, as well as the researchers of other manuscript material, base their research mainly on linguistic elements, while some of them base it mainly on physical features. While proposing different methodologies for those researches, TEI and TEIMark for basically linguistic and DocMark for physically oriented researches, our intention was to avoid the transcription of texts in cases when the text itself was not the focus of the research, but rather physical and other visual features of the manuscript. Different tools will be proposed for those sometimes interwoven but mostly different kind of researches.

In his article on challenges, perspectives and questions imposed during exploration of manuscript collections in the countries of Eastern and South-Eastern Europe, Erich Renhart stresses scientific aspects of those collections that should be examined and points out the importance of interdisciplinary collaboration in the statement that epigraphy, calligraphy, codicology, palaeography, bibliography, linguistics, art history, history, and other fields of study equally meet on the same object – the manuscript.[5] Although interdisciplinary collaboration is well-known in the field of manuscript studies, it has started to flourish when digital technologies entered the field and enabled projects based on collaboration of various researchers. Inman, Reed and Sands described it in the book on the collaboration in the humanities: "Interdisciplinary aided by the digital technologies seems to be a hallmark of the contemporary humanities, and consequently, collaborative projects gathering scientists from various fields of study have become increasingly prominent in the humanities in recent years".[6] While proposing methodologies for manuscript research in this article, we had in mind the necessity of collaboration of various researchers and scientists on the same manuscript. Therefore, all of the methodologies are based on providing appropriate virtual space for collaboration.

5 See: Renhart, Erich. Tracing our written heritage : challenges, perspectives, questions. // Summer School in the Study of Old Books, Zadar, Croatia, 28 September to 2 October 2009 : proceedings / edited by Mirna Willer and Marijana Tomić. Zadar : Sveučilište, 2010. P. 107.

6 Inman, A. James; Cheryl Reed; Peter Sands. Preface: Issues and options for electronic collaboration in the humanities : a framework. // Electronic collaboration in the humanities : issues and options / edited by James A. Inman, Cheryl Reed, and Peter Sands. Mahwah : Lawrence Erlbaum Associates, 2004. P. XVIII.

Annotation

Annotation is generally referred to as "being the process of adding notes to a text or diagram giving explanation or comment.[7] A more specific definition of annotation in our context is the one by the World Wide Web Consortium's (W3C) Annotea project, which states that "By annotations we mean comments, notes, explanations, or other types of external remarks that can be attached to any Web document or a selected part of the document without actually needing to touch the document".[8] This definition, however, should be handled cautiously as there is still an open debate about the issue of whether annotations should be stored within a digitised document or remotely, as it is being suggested by the Annotea team.[9]

Texts are more than sequences of encoded glyphs, because they have structure, content and multiple readings. Encoding or markup is a way of making specific features implicit to a person, explicit to a machine. The growing amount of information that is associated with the text by means of markup requires structuring which does not necessarily result in identical combinations or groupings of units of information. Andreas Witt and Dieter Metzing[10] distinguish between annotation level referring to the conceptual level of information represented in markup, and annotation layer referring to the technical realisation of markup.

The term *level* refers to a model involving theoretical concepts of a specific research discipline. In linguistics, there are several sub-disciplines which investigate different aspects and modalities of natural language and natural language description such as phonology, morphology, syntax, and semantics, which are often called the linguistic levels of description. Thus, an annotation unit (an XML element or attribute) will refer to one level while another annotation unit may refer to another level of linguistic description. In that sense, different levels of markup can be found in one annotated text. But even at one linguistic description level, different types of analyses can be represented which we still consider as different conceptual levels of markup. The

7 Oxford Dictionaries. Concise Oxford English Dictionary. 11th ed. Oxford : University Press, 2008. P. 53.

8 Annotea project : overview [cited: 2011-12-01]. Available at: http://www.w3.org/2001/Annotea/

9 Ibid.

10 Linguistic modeling of information and markup languages / editors Andreas Witt, Dieter Metzing. Heidelberg [etc.] : Springer, 2000. Pp. 1-22.

term *layer*, on the other hand, refers to the technical realisation of a modelling task. What it means exactly depends thus on the annotation system employed: in transcription systems based on the annotation graph framework,[11] for example, a layer corresponds to a single labelled path which spans the transcribed text. Another example of a technical realisation is the use of different XML documents to store annotations of one text, each XML file then corresponds to one annotation layer.

In this article we will concentrate only on the *layer* level, which is technology, so we will name the setting of markers, which are called tags in XML-tagging, in order to differentiate it from the process of *level* marking, which we usually call *annotation*. This has been underlined in the title of the article.

XML (eXtensible Markup Language) as TEI container

XML is a specification for storing information and for describing the structure and meaning (semantics) of that information.[12] It is a meta-markup language that defines a syntax in which other domain-specific markup languages (like TEI, Text Encoding Initiative[13]) can be written. The XML specification enables users to define their own markup language. The only condition is that these newly created tags adhere to the rules of the XML specification.

An XML element is the most basic unit of XML document. It can contain text, attributes, and other elements. A typical XML element is comprised of an opening tag (which can contain attributes with their associated values), content, and a closing tag. The example (with arbitrary names and values) is shown in Figure 1.

The first 'bibl' element has five nested elements (twice 'title', 'author', 'publisher' and 'date') mixed with text. The first 'title' element has an attribute called 'stage' whose value is 'a1', and content 'The Interesting story...'. The element has an opening tag with a name written between less than (<) and greater than (>) signs. The name of the element should describe its purpose and, in particular, its contents. An element is generally concluded with a closing tag, comprised of the same name

11 Bird, Steven; Mark Lieberman. A formal framework for linguistic annotation. // Speech communication 33, 1-2(2001), 23-60.

12 W3C. Extensible Markup Language (XML) [cited: 2011-12-01]. Available at: http://www.w3.org/XML/

13 TEI: Text Encoding Initiative [cited: 2011-12-01]. Available at: http://www.tei-c.org/index.xml

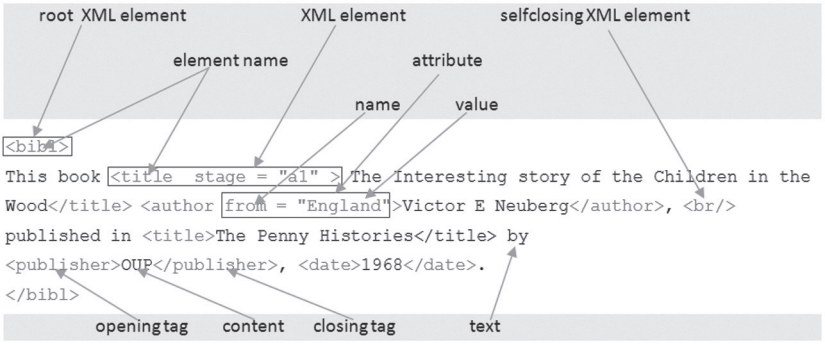

Figure 1.
XML notation

preceded with a forward slash, enclosed in the less than and greater than signs. The exception to this is called an empty element which may be "self-closing".

Elements may have attributes. Attributes, which are contained within an element's opening tag, have quotation-mark delimited values that further describe the purpose and content (if any) of the particular element. Information contained in an attribute is generally considered to be metadata; that is, information about the data in the element (e.g. the author is from England, in the example above), as opposed to the data (e.g. 'Victor E. Neuberg') itself. An element can have as many attributes as desired, as long as each has a unique name.

XML has a structure that is extremely regular and predictable. It is defined by more than 100 different rules, although only several of them are significant (i.e., the root XML element is required, every element must have a closing tag, elements must be properly nested, elements' and attributes' names are case sensitive and an attribute's value must always be enclosed in quotation marks). If an XML document satisfies these rules, it is considered well-formed. Once a document is considered to be well-formed, it can be used in many different ways.

XML predefines no elements at all, but the documents built from them are not completely arbitrary. They must be valid – their structure has to meet some other rules. Good forming is the minimum criterion necessary for XML processors and browsers to read XML documents (files), so these rules are designed to be understood by software rather than human beings. There are two ways to define the structure of an

XML document - either written with a DTD (Document Type Definition) or with the XML Schema language. These structural definitions (or schemas) specify: the name of the root element, names of all elements used, names and data types (strings, date, number etc.) and (occasionally) default values for their attributes, rules about how elements can nest, and a few other things, depending on the schema language. A schema does not specify anything about what XML elements are, what their content "means".

TEI provides a framework for the definition of more than 500 useful textual distinctions (XML elements with their associated attributes), and also provides a set of modules that can be used to define schemes making those distinctions. These schemes ensure that:

- TEI documents use only predefined elements, attributes and entities,
- The same thing is always called by the same name,
- TEI documents are well-formed and valid (enforcing structural rules).

For TEI users the most popular XML schema is Relax NG, although there are many TEI documents verified by simple DTD scheme, which is nowadays outdated.

Since XML tags (and TEI too) are created from scratch, those tags have no inherent formatting and the browser cannot know how to display them. Therefore, it is the user's job to specify how an XML/TEI document should be displayed using general purpose programming languages (like Python, Java, Ruby) or using XML-markup language named XSL, or eXtensible Stylesheet Language. XSL is actually made up of three languages: XSLT, for transforming XML documents; XPath, for identifying different parts of an XML document; and XSL-FO, for formatting an XML document. XSL lets one manipulate the information in an XML document into any format one needs; most frequently into HTML, or an XML document with a different structure than the original (Figure 2).

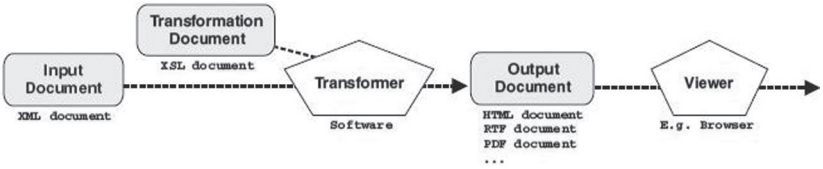

Figure 2.
Transformer software

As it might be seen from the diagram, we need the Transformer - software which abiding by certain rules will turn our XML document into an output document of different format (e.g. HTML or PDF) which we might be able to browse or view in a different way, with a special browser/viewer.

Proposed methodologies for manuscript research

Word markup procedure using TEI (Text Encoding Initiative)

In their study of ancient books, the book historians and philologists often need to markup, count, compare or perform some similar action on some part of a text or a text component such as personal names, places, foreign words or even linguistic categories (nouns, verbs, adjectives, etc.). If they did this without using computer, they usually circled or underlined such words in their scientific research on the copies of the original text or they even created the supporting text content as a list of personal names, foreign words etc. Such research data served to confirm or refute the set scientific theories.

However, in using computer for the text markup procedures it is also possible to implement some standard editing programs such as MS Word or Open Office. For the markup of a computer-readable text, like a Glagolitic and transliterated version, a common editor's comment function can very well serve the purpose, as shown in the Figure 3:

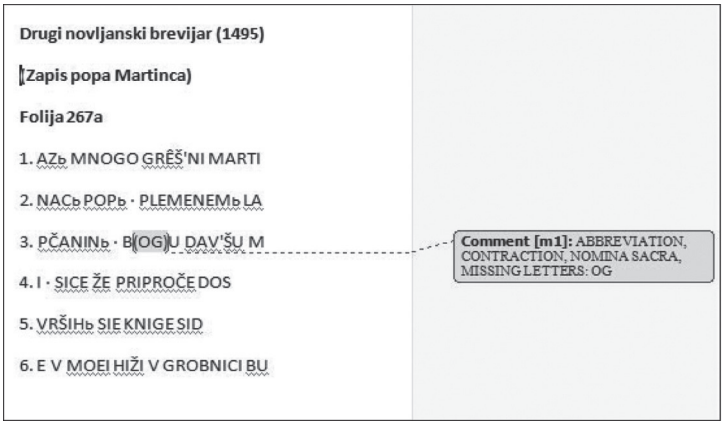

Figure 3.
Encoding of acronyms in a Glagolitic text of the the *Second Novljanski Breviary* using the comment function in MS Word – a word processing program

However, if a scientist wishes to search, scan or even extract the highlighted text, this method of labelling will become insufficient. Word processing programs allow the implementation process of markup without any possibility to elaborate the tags and selected words; therefore, they are similar to manual word selection, without the use of computers. Consequently, other markup procedures have been proposed for word selection using computer, among which the most outstanding are related to the use of tagging (markup) languages and their applications (programs).

TEI (*Text Encoding Initiative*) guidelines are widely used for encoding of texts from human sciences. TEI represents the word encoding standard which has been developing since 1990, and has experienced as many as five versions so far (P1-1990 up to P5- 2007). The TEI specification provides an extremely rich set of elements for almost all conceivable text types (prose, poetry, drama, novel and alike) with appropriate attributes,[14] although in practice, in individual cases, a much smaller number of elements is associated with each text type. The TEI standard modularity is one of its most conspicuous features, because the scientist may adapt it or customize it to his/her own needs. Sometimes it will refer to some ancient book, sometimes to some mediaeval codex manuscript and alike. Depending on the text type, in particular considering the form and content, the TEI standard recommends the use of certain subset of TEI elements. This means that each scientist may freely mark up the text for his/her project but s/he is also required to construct, as we have mentioned, his/her own scheme document (DTD, XML Scheme or RELAX NG) where s/he shall define the selected subset of elements from the TEI specification as well as their application in the text markup procedure.[15] One among the most popular TEI standard-customizations is *TEI-Lite*, first published in 1995, and currently consisting of (only) 145 elements, but it meets the requirements of 90 % text markup projects within the TEI community.[16] The TEI-Lite is also the first step in the implementation of TEI markup texts. Since each and every XML is expandable, whereas the TEI happens to be only an XML variant, it is possible to add into the TEI pro-

14 Accurate number of TEI elements and attributes in the TEI standard (at the beginning of 2011) is 503 elements and 210 attributes.

15 In this sense we say that the user performs some customization or personalization of the TEI standard to his/her own needs.

16 TEI by example : tutorials : introduction to text encoding and the TEI [cited: 2011-12-01]. Available at: http://tbe.kantl.be/TBE/modules/TBED00v00.htm

gram some completely new, one's own elements, along with the already selected set of TEI or TEI-Lite elements, as well as to introduce elements from any other set of elements, such as MARC 21 or MathML. To customize the TEI set of elements to one's own needs, from the P5 version onwards, it is possible to use the special network software – *Roma*.[17] In the Roma interface (Figure 4) it is possible to choose exactly those elements and attributes which meet the specific requirements of a text encoding, and which would finally produce (generate) the associated scheme document.

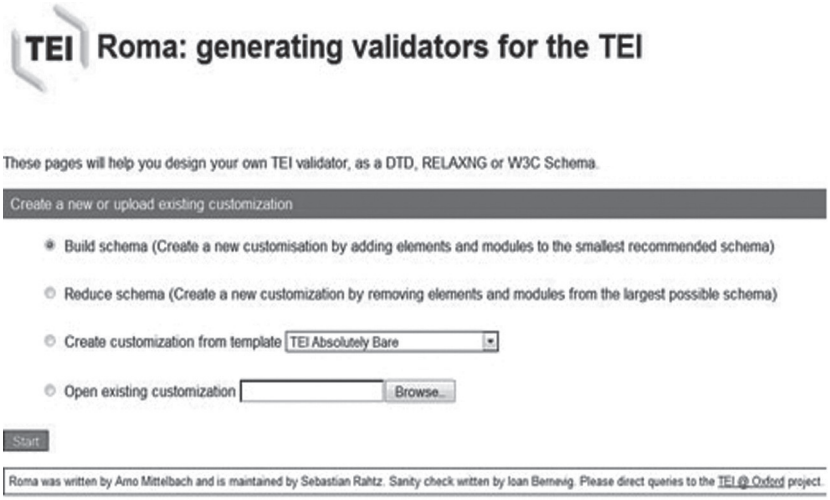

Figure 4.
Roma, the network software for the TEI customization to one's own needs

Having generated the appropriate scheme document, it is necessary to select the XML editor for the word markup procedure. Apart from the text which needs to be marked, editor will also load the generated scheme, which will cover all allowed elements and attributes (as to where and how) implemented by the user in the text markup procedure; to be more precise, it will control and manage the text encoding. Nowadays there are numerous XML editors on the market, among which the best-known are: Oxygen XML Editor[18] and Altova XML-

17 Roma : generating validators for the TEI [cited: 2011-12-01]. Available at: http://www.tei-c.org/Roma/

18 <Oxygen/> XML Editor [cited: 2011-12-01]. Available at: http://www.oxygenxml.com/

Spy[19] as a commercial, and PSPad[20] and Notepad++[21] as free (open-source) solutions. For example, with the Oxygen XML Editor the text markup procedure is highly facilitated due to the possible choice of elements and attributes, as determined by the previously loaded scheme document (Figure 5).

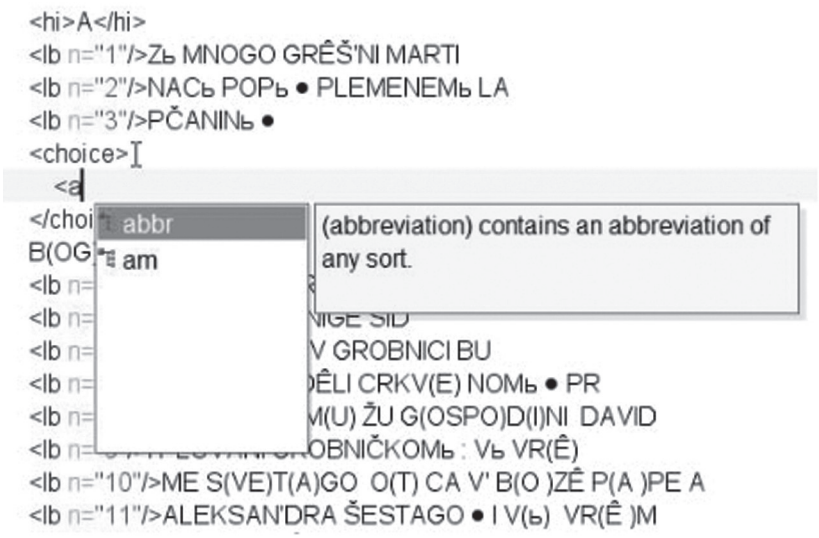

Figure 5.
Implementation of the text markup procedure using *Oxygen XML Editor*

The TEI annotated text needs to be processed or elaborated if it is to provide the scientist with the quantitative research data (e.g. frequency of certain elements in the text) or lists of the selected data types and alike. Figure 6 shows the list of all words (in colour) containing the grapheme 'nj' obtained using the XSLT program. In this way it is possible to count the selected elements, to list them, compare them, etc., by means of the XSLT program's XPath and XSL functions.

19 Altova XMLSpy : XML editor for modeling, editing, transforming, & debugging XML technologies [cited: 2011-12-01]. Available at: http://www.altova.com/xmlspy.html
20 PSPad [cited: 2011-12-01]. Available at: http://www.pspad.com/
21 Notepad++ [cited: 2011-12-01]. Available at: http://notepad-plus-plus.org/

U	Piſanju	i	ſtampanju	ricih
zamerſsenja	ima	tako	da	nike
nego	druge	i	tako	smetnja
slova	na	izuſschivanje	ricſih	metati
kolikobi	u	naſs	iezik	potribito
na	izuſschivanje	naſsih	ricſih	manjkaju
nemore	dochi	na	iedinoſt	itako
ſva	kolika	ta	mucſnoſt	dolazi
naſsi	Alli	ier	nadopuniti	po
pomochna	slova	neſtoie	kod	dvoice

Ukupan broj označenih grafema: 66

Ukupan broj **grafema** *'nj'*: 17

Piſanju

ſtampanju

zamerſsenja

Figure 6.
With XSLT it is possible to display the selected words in another colour by means of TEI

The text from the field of humanities marked by the TEI standard provides the researcher with the opportunity to analyse words not only quantitatively but also qualitatively, e.g. to generate the controlled dictionary of names based on selected names from text. However, the processing of the selected words using XSLT or any other program language (Java, Python or Ruby) requires additional knowledge and programming skills for every single application. Unfortunately, a general TEI tool for all kinds of tasks that are set before the scientists who want to process each digitised document or a transcribed manuscript has not been developed yet.

TEIMark: a web-based editor

TEIMark is the conceptual solution of a web-based editor for encoding text using TEI standard. Web-based means that application is accessible over the Internet and the only thing needed to use it is a web browser. TEIMark is developed using jQuery and JavaScript technologies on the client side and PHP programming on the server side. The base of the project is TinyMCE visual editor which is expanded to support TEI XML tags and attributes. TinyMCE is an open-source, platform independent WYSIWYG (what you see is what you get) editor (Figure 7).

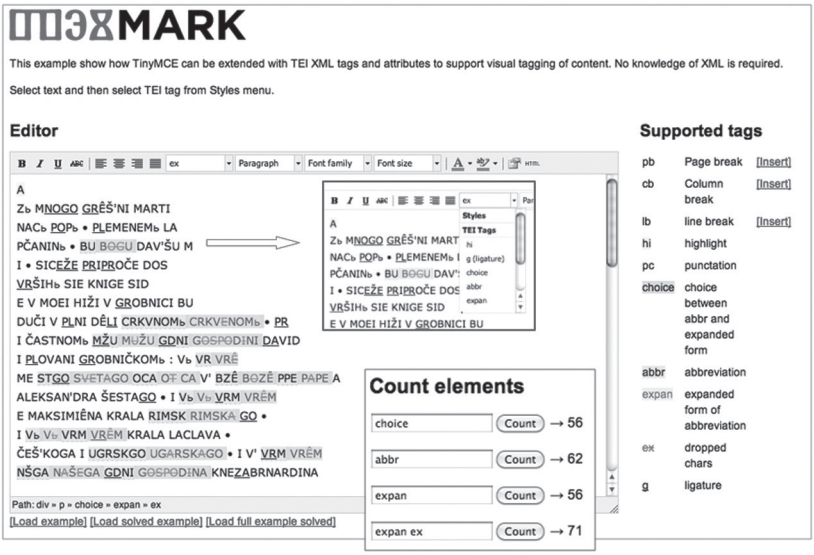

Figure 7.
Application workflow demonstrated with a 10-line example from the the *Second Novljanski Breviary.*

Encoding is done by selecting the words, characters or symbols we want to markup and selecting a proper TEI tag from styles dropdown menu (so virtually no XML knowledge is necessary). Tags are made visible by using different colours and type treatments (legend for supported tags is shown next to the editor). The nested tags are also supported. Depending on the chosen tag there is a visual distinction between different tags. After encoding is done the output can be saved as a standard TEI XML file.

The second part of the application is the ability to edit TEI XML header. This is done in different text fields because visual editor is used only for content while headers and other metadata are kept logically separated. In this part of application the complete TEI XML file, including headers and content can be downloaded, too. The third part of the TEIMark is an interface for counting elements that are visually encoded using editor. Counting is done by jQuery selectors so it is easy to count nested elements or elements with specified attributes (e.g. for number of dropped characters within expanded form of abbreviation *expan ex* is put into the field). Any marked up element like punctuations, ligatures or ligatures within abbreviations can be counted. Some Glagolitic web fonts are also included, so they can be chosen from the font menu to display the text in Glagolitic font.

The TEIMark program allows the users to: load ordinary or transliterated text into editor; make selections from words, characters and symbols and encode them with TEI tags using drop-down menu; insert page, column and line breaks using legend of supported tags; edit raw TEI XML header; download final TEI XML file which includes header and encoded content; analyse and count encoded elements by writing single tag or combination of tags; and select text and display it in one of two Glagolitic typefaces.

DocMark: software for visual encoding of digitised images

As indicated by its name, TEI (Text Encoding Initiative) is the most appropriate for text based researches. But, in a wide range of palaeographic and codicological researches, as well as in those in the field of art history etc., it is not the text that is in the centre of a research, but its physical context which is hidden when one works on a transcribed version of a document. It is, for example, impossible to research a *mise en page* or to distinguish among different hands in the same manuscript if a research is conducted on the transcribed text, and not on the image of it. Following that approach, initiated by Elena Pierazzo and Peter A. Stokes and their idea of "putting the text back into context",[22] we are proposing a tool for the research of the digitised manuscripts. It allows a research on features that are not visible in the transcript of a document, like a change of morphology of a letter made by different scribes or any notes made in the rich history of researched manuscripts, the bindings, the holes in the parchment etc.

22 See more on the subject in Stokes, P. A.; E. Pierazzo. Op. cit., pp. 397-430.

Since the 1950s digital imaging has begun to revolutionise manuscript research by providing broader and lower-cost access to higher-quality manuscript images, and by facilitating the research through improving legibility of damaged text etc.[23] As Fischer and Sahle point out that "digital facsimiles convey a great number of original features and characteristics and can be easily provided and shared",[24] it is undoubtedly recognised among scholars that research in palaeography, codicology, art history, history of the book and other disciplines can be greatly facilitated by means of manuscript facsimiles provided along with codicological data and descriptive texts.

Software that enables such an approach to encoding and processing of digitised documents is DocMark. It provides encoding of digitised image by setting tags directly to the digital facsimile. In this way it becomes possible to encode target elements on the document without losing of a number of material properties of the document, otherwise lost in the transliterated text, to set a number of the same tags on the documents written in different languages and/or scripts and to later analyse (of individual elements, their comparisons, etc.), as well as to browse the documents' base by the encoded elements.

DocMark, a web based computer system, is using new WebGL/Javascript/HTML5 and Ruby/SQLite technology. Before encoding, the author assigns some feature/category (analogue to selection of some TEI tag) to every tag. It is also possible to add a special description to every single tag within the same type. Upon tag selection, the user sets the desired visual tag by clicking onto a desired place for a document image. There are several tag types (e.g. a triangle, a square, an arrow, a cross etc.) and several hundred pre-assigned symbols (obtained from the right-click menu trees) in different colours, sizes and transparencies (Figure 8).

DocMark also offers a kind of a collaboration platform which enables researchers to work at one or several layers over the same document image, either locally or remotely through the network.

Precise measurement and marking of particular parts of a document is also possible. Among tags there are also lines which can be

23 Griffin, Carl W. Digital imaging : looking toward the future of manuscript research. // Currents in Biblical research 5, 1(October 2006), 58-72.

24 Fischer, Franz; Patrick Sahle. Introduction: Into the wide – into the deep : manuscript research in the digital age. // Codicology and paleography in the digital age 2 / edited by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt : Herstellung und Verlag, 2010. P. XI.

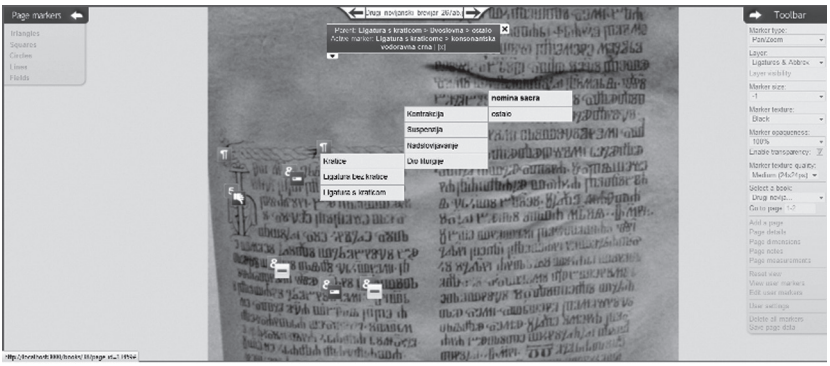

Figure 8.
Visual encoding of digitised page of the the *Second Novljanski Breviary* using DocMark

drawn by the researcher from one position to another in order to encode the researched area in the document. All tags, measurement and descriptions are stored in a database and are retrievable. Documents can be searched by all encoded features, tags can be analysed and compared on one or several documents, while document can be viewed with all or some specific tags, which enables a complete visualisation of encoded features (Figure 9).

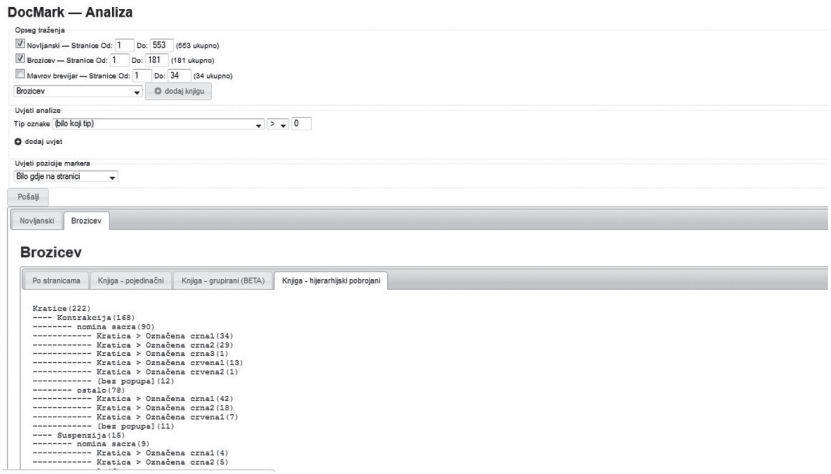

Figure 9.
An analysis of material encoded using DocMark

There are similar projects, yet in terms of conception and technology they are different: TILE,[25] Image Markup Tool,[26] and Edition Production & Presentation Technology (EPPT)[27] facilitate document tagging, but the result is an XML record. In the case of DocMark, the original image of a document is preserved, while layers with relevant information on all set tags with their accurate positions are stored in a computer database.

History Integrator

As the text is not only a sequence of encoded glyphs, the document/manuscript, too, is not exclusively made up of sequences of files stored in a computer database, but is part of a culture, space and time in which it occurred. To visualise this important segment of the manuscript, the authors of DocMarc have developed a computer program called History Integrator. It connects documents by the time of their origin, the place where they were created or are currently kept, and by the information they contain. The time axis is made by the MIT Simile project and the area is covered by Google Maps.

The encoding of digitised documents has been presented in both variants so far: as transcribed (digitised) text and as an image of the manuscript. For the first one, editor (TEImark) has been proposed whereby the user would not have to be familiar with the XML-technology, but only with the TEI classification and its labels, whereas for the second one – visual encoding – setting up of visual tags/markers in layers above the document has been proposed. All that remains is the inclusion of documents into the history, in terms of time and space (spatial and temporal coordinates). It is also a kind of document indexing, but not any more of letters, words or characters they consist of, but of the documents themselves. Moreover, this new designation into the time and space scale (at a country, city, or even building level, in Google map perspective) provides linking with the digitised document. The user chooses an author who was entered into the timeline scale (e.g. Blaž Baromić from 1450); then s/he obtains basic information on him,

25 TILE: Text-Image Linking Environment [cited: 2011-12-01]. Available at: http://mith.umd.edu/tile/

26 The UVic Image Markup Tool Project [cited: 2011-12-01]. Available at: http://tapor.uvic.ca/~mholmes/image_markup/

27 Edition Production & Presentation Technology (EPPT) [cited: 2011-12-01]. Available at: http://eppt.org/eppt/

a list of his works, and if the works have been digitised and saved, then s/he gets the corresponding links, too. By clicking on such a link s/he may read, browse and search the document by tags entered through the DocMark, for each and every page - this is the key integration property of the History Integrator and the DocMark program.

Each user in a collaborative project can contribute his/her own content to each and every data element (author, document, event and alike) published in the History Integrator. In this way a new kind of visual encyclopaedia can be created, because apart from the published material, descriptions, explanations and supplementary details are added, too.

History Integrator is a program currently used for development and testing of new ideas in the field of humanities. This refers in particular to the concept of information with its contents, place of origin, and the route it made, its current storage, time of its creation or any subsequent modification. Information may refer to people, events and/or their work. So the route made by a certain book, place and fragment contents, authors' biographies can be recorded, but the digitised works by the same author may be also retrieved, even those tagged by the agreed metatags (either by TEIMark or by DocMark) (Figure 10).

The program provides for the setting, modification and deletion of information entered about author, event or work, as well as interconnection of data elements from different users into desired units.

Every user disposes of a time and space axis (taken from MIT open-source project) onto which h/she enters his/her particular data (event or term). These terms represent categories which have to be hierarchically placed one upon another, for the sake of navigation and availability. The categories conform to a thesaurus structure (a tree) and may be easily adapted to future ontological records (graphs). In defining of a new term, the user shall use the previously built up structure, so s/he shall search and find the higher term (super-category) to which the particular term actually belongs. If a higher category is not available, the word becomes a new root (source) of future sub-terms. The result is a qualitative categorisation (classification) of information. Every word may, by thesaurus' rules, have one or more synonyms and one or more similar words.

The terms are described by additional data which are entered into a computer database by means of an online editor or can be retrieved from a web site. The described programs include all details needed for

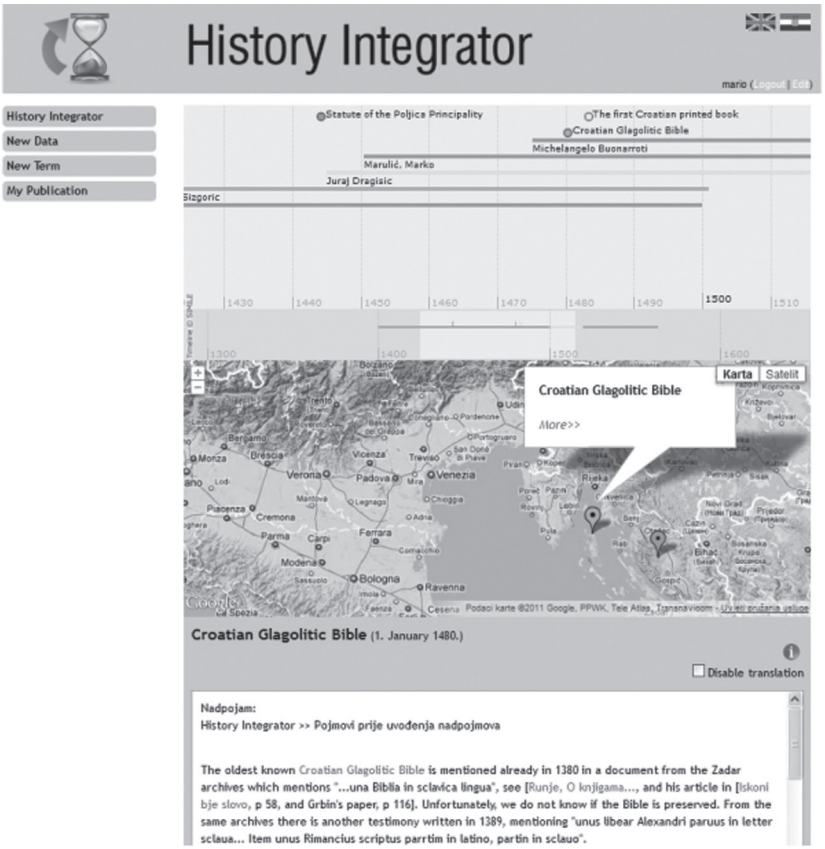

Figure 10.
History Integrator

RDF (Resource Description Framework) publishing in linked data global cloud, but this is part of another story.

If the information content disposes of a term which has been already entered into the database, the term will be underlined, so the user may (if s/he places the mouse cursor above the underlined word or a phrase) get a window with the description of this term. The structure is recursive, therefore in the data of the newly created window all familiar terms have been underlined, too, which provides for the opening of new (sub)windows and data cohesion in all directions (over synonyms and similar terms). In this way it is possible to perform automated tagging of terms in information content where XML-tags are set around

the found term by the predefined scheme. This has largely accelerated the tagging of documents; major work has been done by computer, however, manually performed, individual interventions always remain possible. This is the first step in a quick data preparation, for a cloud computing - linked data in the semantic web.

Being a web application it allows collaboration among users which has been developed in order to avoid repetition or redundancy, such as entering the already recorded information. Moreover, the program History Integrator provides each and every user with a possibility to pick from other users a piece of information (only the one which has been offered, shared and/or allowed) and place it on their time and space axis or on their tree of terms. In the same way, they can offer their own work to other users, to the extent they deem appropriate. In case that several users have coded the same information and offered it to others, the user may select only one piece of information. In this way the effectiveness of cohesion has been solved and redundant information reduced.

Conclusion

In this article, different methodologies of manuscript research in digital environment based on annotation have been described. For two different types of researches, those based on a text, and those based on a physical context, different types of annotation methodologies are proposed: TEI and TEIMark that serve for encoding of transcribed digitised manuscripts, and DocMark for encoding and analysing of digitised versions of manuscripts. Additionally, History Integrator software which places manuscripts into the time and space framework with additional multimedia information is described. The common characteristic of the described programs is that they have been developed under the supervision of humanities' scientists, and that they have not been run on local computers but online, through Internet. Such an option provides many scientists from different branches of humanities with the possibility to jointly collaborate on the same manuscript, regardless of a computer or operating system used, since they need only one program for their work – a web browser. Ensuring platform for virtual collaboration of remote scientists working on the same manuscript is also considered as one of prerequisites for manuscript research in the digital age which is enabled both by TEI based and image tagging based software.

References

Altova XMLSpy : XML editor for modeling, editing, transforming, &debugging XML technologies [cited: 2011-12-01]. Available at: http://www.altova.com/xmlspy.html

Annotea project : overview [cited: 2011-12-01]. Available at: http://www.w3.org/2001/Annotea/

Bird, Steven; Mark Lieberman. A formal framework for linguistic annotation. // Speech communication 33, 1-2(2001), 23-60.

Edition Production & Presentation Technology (EPPT) [cited: 2011-12-01]. Available at: http://eppt.org/eppt/

Fischer, Franz; Patrick Sahle. Introduction: Into the wide – into the deep : manuscript research in the digital age. // Codicology and paleography in the digital age 2 / edited by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt : Herstellung und Verlag, 2010. Pp. XI-XVI.

Griffin, Carl W. Digital imaging : looking toward the future of manuscript research. // Currents in Biblical research 5, 1(October 2006), 58-72.

Hercigonja, Eduard. Trojezična i tropismena kultura hrvatskog srednjovjekovlja. 2. dopunjeno i izmijenjeno izd. Zagreb: Matica hrvatska, 2006.

Inman, A. James; Cheryl Reed; Peter Sands. Preface: Issues and options for electronic collaboration in the humanities : a framework. // Electronic collaboration in the humanities : issues and options / edited by James A. Inman, Cheryl Reed, and Peter Sands. Mahwah : Lawrence Erlbaum Associates, 2004. Pp. XVII-XX.

Linguistic modeling of information and markup languages / editors Andreas Witt, Dieter Metzing. Heidelberg [etc.] : Springer, 2000.

Notepad++ [cited: 2011-12-01]. Available at: http://notepad-plus-plus.org/

Oxford dictionaries. Concise Oxford English dictionary. 11th ed. Oxford: University Press, 2008.

<Oxygen/> XML Editor [cited: 2011-12-01]. Available at: http://www.oxygenxml.com/

PSPad [cited: 2011-12-01]. Available at: http://www.pspad.com/

Renhart, Erich. Tracing our written heritage : challenges, perspectives, questions. // Summer School in the Study of Old Books, Zadar, Croatia, 28 September to 2 October 2009 : proceedings / edited by Mirna Willer and Marijana Tomić. Zadar : Sveučilište, 2010. Pp. 107-118.

Roma : generating validators for the TEI [cited: 2011-12-01]. Available at: http://www.tei-c.org/Roma/

Stokes, Peter A.; Elena Pierazzo. Putting the text back into context : a codicological approach to manuscript transcription. // Codicology and paleography in the digital age 2 / edited by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt : Herstellung und Verlag, 2010. Pp. 397-430.

TEI: Text Encoding Initiative [cited: 2011-12-01]. Available at: http://www.tei-c.org/index.xml

TEI by example : tutorials: introduction to text encoding and the TEI [cited: 2011-12-01]. Available at: http://tbe.kantl.be/TBE/modules/TBED00v00.htm

The UVic Image Markup Tool Project [cited: 2011-12-01]. Available at: http://tapor.uvic.ca/~mholmes/image_markup/

TILE: Text-Image Linking Environment [cited: 2011-12-01]. Available at: http://mith.umd.edu/tile/

W3C. Extensible Markup Language (XML) [cited: 2011-12-01]. Available at: http://www.w3.org/XML/

Žagar, Mateo. Osnovni procesi konstituiranja ustavne glagoljice. // Вьlgari i Ньrvati prez vekovete II / ur. Rumjana Božilova. Sofija : IK Gutenberg, 2003. Pp. 31-42.

Biographical sketch

Dr. Mario Essert is a full professor at the Department of Control Engineering at the Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia. He received his PhD in the computer science from the Faculty of Electrical Engineering of the University of Zagreb. As a member of the Group of Discrete Mathematics at that Faculty (led by Prof. V. Ćepulić) he participated in four international co-projects in Mainz (Germany), Kiev (Ukraine), Hangzhou (China) and Heidelberg (Germany). His research interests include combinatorial algorithms, computer mathematics, distance learning and web programming.

Dr. Boris Bosančić is a research assistant at the Department of Information Sciences at the Faculty of Philosophy, J. J. Strossmayer University of Osijek, Croatia. The aim of his PhD thesis was to examine the usefulness of text encoding using TEI standard in digital library environment for the researchers in social science and humanities who study old Croatian books. He is associate on the project which is financed by the Croatian Ministry of Science, Education and Sport *Digital Library of Croatian Printed Heritage by 1800: Structural Premises.* His research and professional interests are markup languages (XML, RDF/XML), metadata and identifiers, linked data, ontology (OWL), and text encoding (TEI).

MarijanaTomić is a research assistant at the Department of Library and Information Sciences at the University of Zadar, Croatia. She is enrolled in the PhD study programme of mediaeval science at the Faculty of Humanities and Social Sciences, University of Zagreb, Croatia. She is one of the coordinators of the project of organising the library of the monastery of St. Francis in Zadar, and an associate on the project *Digital Library of Croatian Printed Heritage by 1800: Structural Premises* which is financed by the Croatian Ministry of Science, Education and Sport. Her research interests are information organisation, cataloguing of old and rare books, history of the book, digital humanities, mon-

astery libraries, Glagolitic script, organisation and appropriation of Croatian mediaeval Glagolitic texts.

Mario Lončarić is a student at the Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia. He works as a lead web developer at a full service digital agency that offers creative solutions in the field of digital media. His professional interests include content management systems, database design, programming and developing user interfaces with strong accent on usability and accessibility.

OD XML-A DO VIZUALNOG OZNAČIVANJA PRIJEDLOZI METODOLOGIJA ZA ISTRAŽIVANJE DIGITALIZIRANE SREDNJOVJEKOVNE GLAGOLJIČKE GRAĐE

Sažetak

Rukopisi su često temeljni istraživački izvori za istraživače iz raznih disciplina. U današnje vrijeme ta su istraživanja potpomognuta novim alatima temeljenim na informacijskoj tehnologiji. U ovom članku opisane su različite metodologije temeljene na označivanju koje se mogu koristiti pri istraživanju rukopisa u digitalnom okruženju. U prvom dijelu članka objašnjeni su pojmovi označivanja i XML-a (eXtensible Markup Language). U drugom su dijelu prikazane rezličite metodologije temeljene na označivanju koje se mogu primijeniti u dvama vrstama istraživanja, i to TEI i TEIMark koji se koriste za označivanje transkribiranih inačica rukopisa i preporučuju se za istraživanja samih tekstova, i DocMark koji se koristi za označivanje i analizu digitaliziranih inačica rukopisa i preporuča se za istraživanja teksta u ukupnošću s njegovim fizičkim kontekstom. Uz to, opisan je History Integrator, program koji rukopise smješta u kontekst vremena i prostora te im nadodaje multimedijski sadržaj.

Ključne riječi: digitalna humanistika, označivanje, XML, TEI, alat za označivanje slika, analiza slika označenih u mrežnom okruženju, hrvatska srednjovjekovna glagoljička građa