Digitisation of Literary Heritage Using Open Standards

Tomaž ERJAVEC[1], Matija OGRIN[2]

[1] *Department of Knowledge Technologies,*
Jožef Stefan Institute
Jamova 39, Ljubljana, SI-1000, Slovenia
tomaz.erjavec@ijs.si

[2] *Institute of Slovenian Literature and Literary Sciences,*
Scientific Research Centre of the Slovenian Academy of Sciences and Arts
Gosposka 13, SI-1000 Ljubljana, Slovenia
matija.ogrin@zrc-sazu.si

Abstract: The paper presents the methodology, technology and results of a collaborative Slovenian project aimed at e-publishing text-critical editions of literary heritage. The materials exhibit great complexity, as they are made available not only in facsimile but also in several interconnected transcriptions, and can include notes, glossaries, dictionaries, links to external resources, multimedia presentations, etc. Their preparation centres on up-translating the materials into a canonical, standardised edition employing XML and the Text Encoding Initiative Guidelines, and the down-translation of this storage format into the HTML Web presentation. This workflow relies on the use of open standards and intense collaboration between the content and technology providers. We also present the e-editions currently available from the project Web site and discuss further work, esp. the introduction of language technology into the publication process.

1. Introduction

Texts of various kinds – religious, literary, historical etc. – serve as the medium of our cultural memory. Without texts important for European culture we would not know where we come from and who we are: we would miss an essential part of our proper self. Yet, such texts cannot fulfil this mission unless they are understood. And for historical reasons we often cannot understand them in their original form because of the archaic language, old scripts or simply because most readers cannot read (old) handwriting. Overcoming these difficulties is the task of textual criticism and editorial technique, often with the help of ancillary historical disciplines such as diplomatics (codicology) and palaeography.

To make (old) texts comprehensible we first need a critical edition (so called *editio maior*), where the texts are meticulously transcribed from primary sources, if necessary reconstructed, and the text commented. On the basis of such an edition, later simple commercial editions are published (*editio minor*). But critical editions, which contain facsimiles, transcriptions, apparatus and (if necessary) translations into modern languages, are faced with considerable economic barriers, particularly in countries with a small book market, such as Slovenia. For critical editions of older Slovenian texts, *digital* editions are an ideal solution not only because of a better cost–benefit ratio, but also because of the

possibilities that the digital medium offers for the reconstruction, analysis and representation of texts.

The particular methodology needed for this purpose is the main topic of this paper. The methodology should, on the one hand, encompass the specific editorial problems of Slovenian literature and, on the other, be based on open international standards and guidelines for text encoding and interchange. Arriving at this combination is the goal of the collaborative project "Scholarly digital editions of Slovenian literature" (http://nl.ijs.si/e-zrc/). In this paper we present the methodology developed in the first stage of our project, and the main results to date. Hopefully, the first editions reflect the basic tenet of our project: to apply the traditional text-critical and editorial standards, suitable for early Slovenian texts, to the electronic medium, i.e. to join the traditional editorial and modern text-encoding standards.

2. Objectives

Important documents of cultural heritage, such as old literary texts, are a particularly suitable object for multimedia presentations. They are relevant: for educational purposes, as well as for assisting research in, say, literary studies, linguistics, history etc. However, first the text (and related audio/video materials) must be analyzed, encoded and presented in a way that brings out its communication potential to the greatest possible extent. From this follow the goals of our project:

- Assembling the texts in their original form (facsimile) with well-researched transcriptions, apparatus and possibly translations, preferably enriched with multimedia presentations (reading of the materials, video enactments), with interlinkage between these views of the texts, and with external links to related resources (references, historical context).
- Encoding the texts in a format that is able to represent the complexities of text-critical editions, is maximally impervious to technological change, portable across computer platforms and applications, well documented and understandable.
- Enabling access to the texts that can reveal and compare their various views and can be tailored to specific needs and user profiles. Of particular importance is the educational aspect, where, to appreciate the historical peculiarity of an old text and to become aware of the development of written language, the pupils or students must be given the opportunity to compare (the transcription of the) original with its translation into modern language. With the inclusion of audio recordings and other multimedia the educational impact becomes even greater.

3. Methodology

Our project has a structure which is by no means rare in humanities computing: one partner has extensive expertise in the science of textual criticism, which had been, however, done mostly in a classical manner with the computer used only as a word-processor, while the other partner's expertise lies in human language technologies, in particular in the compilation of annotated textual corpora. As one partner was predominantly involved in producing the content and validating the result, and the other in implementing the formal structure and converting into and out of it, it was imperative to enable a seamless platform in which to effect the development of each digital edition.

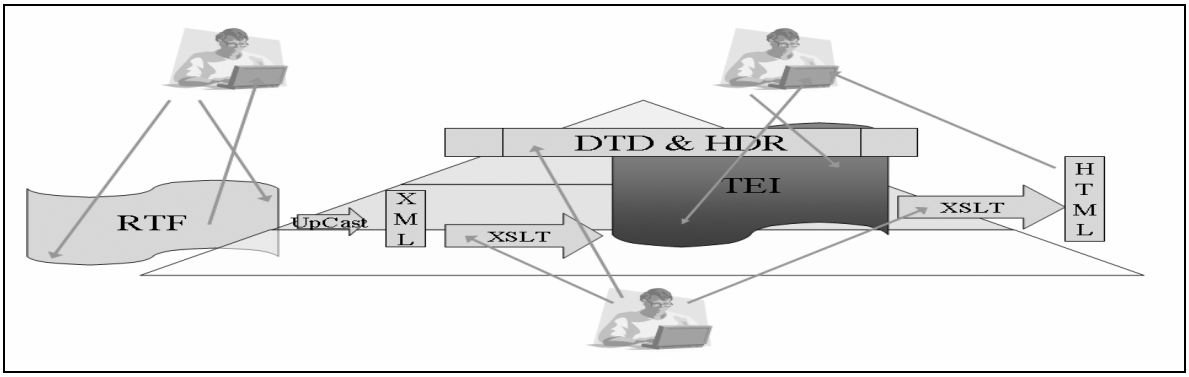

Figure 1:The workflow in the preparation of the materials; the horizontal axis represent time/effort needed to produce a particular resource, while the vertical axis represents the useful information of a resource.

The methodology employed in the production process is illustrated in Figure 1, and centres on the canonical, standardised edition of each material, which is stored in XML, according to a parameterisation of the Text Encoding Initiative Guidelines [1], a specification primarily meant for scholarly encoding of texts (c.f. Section 4). The preparation of the materials revolves around the up-conversion of the original digital document into TEI/XML, and the down-conversion of this storage format into the HTML Web presentation.

3.1 Preparation of the materials

First, the exhaustive text-critical analysis and transcription(s) of text are prepared in a text-editor. As the majority of early Slovenian texts exist in one version (witness) only, the editions generally don't aim to collate versions but to present the autograph itself in its original historical grammar, lexis and orthography. At this stage we must decide which features of the text should be marked-up, i.e. which "intelligence must be embedded in the text in such a way that the program can derive information from it".[2]

Once the analysis and preparation of text is over and the transcriptions, emendations, notes etc. are written in a text editor, usually Word, the material is transformed (with XSLT, c.f. Section 4) into the canonical format. This transformation takes the form of a cyclical process, where the data are, using a dedicated transform, automatically converted into the TEI encoding; this is displayed in HTML using a stylesheet, and the result evaluated. The errors discovered can be one of three types: (1) mistakes in the original file (2) mistakes in the conversion procedure or (3) mistakes in encoding practices. For (1) the original file is corrected, for (2) the transform scenario and for (3) the (semantics of the) XML schema that specifies the element vocabulary used. After correcting the observed mistakes, the up-conversion is re-run on the original file and the cycle repeated. This kind of rapid prototyping approach encourages collaboration and the interchange of expertise.

When the material for a particular edition reaches the stage where the content and annotations reach the limit of what is still possible (and feasible) to correct in the (Word) original, the digital original is discarded, and only the canonical TEI version retained. Any subsequent revisions then proceed directly on the canonical version, using an XML editor. This, of course, means that from this point on the person editing the materials must be familiar with the concepts behind XML and with the TEI encoding scheme. This stage also involves writing the meta-data of the edition, i.e. a description of the material, its sources, and the encoding practices used – these are all contained in the TEI header.

3.2 Presenting the materials

With the canonical, storage format of the materials in place, there remains the question of how best to utilise them. We are, at this stage, not overly concerned with printed versions, but rather aim to offer the materials on the Web, and, possibly, on CD-ROM.

We currently support one HTML view per book, which is produced off-line with an XSLT stylesheet. As can be seen in Figure 2, the HTML itself is reasonably sophisticated – it tries to follow the original as close as possible (structure, placement, emphasis, corrections, emendations), and also provides parallel views of the various transcriptions as well "digital extras", e.g. links to the exact passage of the Bible that is being referred to in the text. We also made efforts for our presentation to conform to relevant accessibility standards [3].

The current scenario of "one book, one HTML" has the advantage of being viewable by any HTML browser, and is suitable for both Web and CD-ROM publication. However, it does not enable the adaptation of the presentation to different needs and user profiles, or take advantage of all the other possibilities offered by the XML annotations. This remains as further work, but it should be noted that software is beginning to be made available that is specifically tailored to enable complex views of TEI encoded text-critical editions [4] – so it makes, at this stage, more sense to concentrate on the content rather than on the mode of presentation.

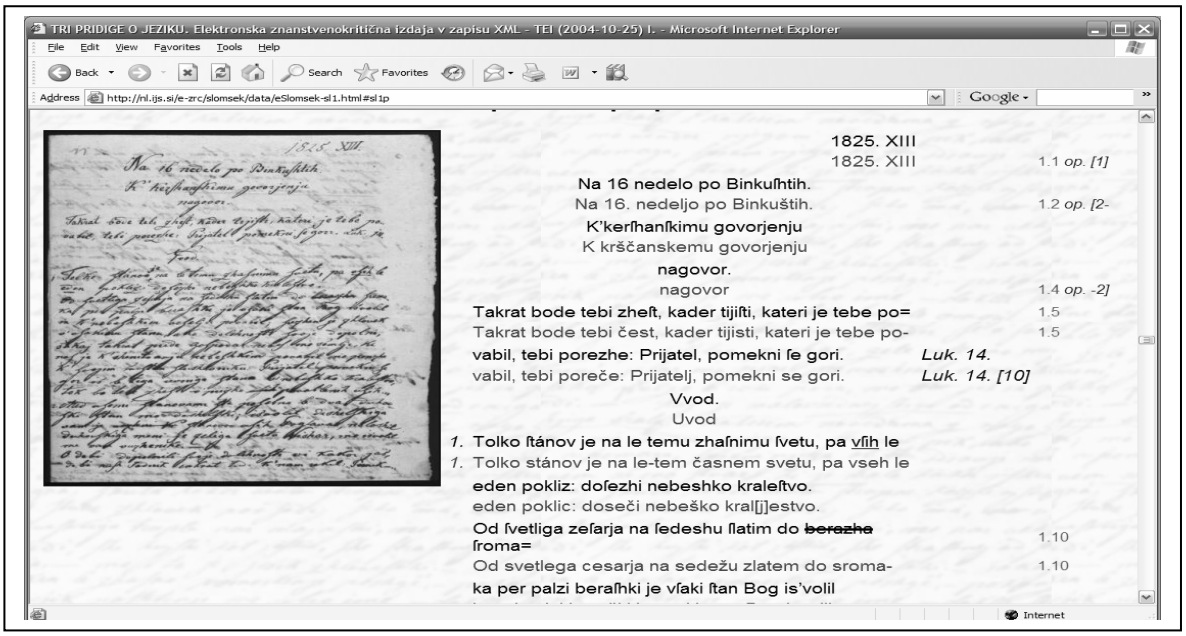

Figure 2: Example of the presentation HTML format

4. Technology Description

The core technology employed in our digital editions is XML. In particular, this involves the definition of the XML element vocabulary (specification of the XML DTD), the construction of up- and down-conversion filters (XSLT), using an XML editor, and the use of special characters that appear in transcriptions of historical texts. In this section we detail these aspects of the technology used in preparing and presenting the materials.

4.1 The TEI Document Type Definition

An XML schema defines the element vocabulary for a particular type of documents and the allowed interrelationships between these elements. It is a vital part of using XML, esp. in the production stage, as it gives the semantics of elements used, and enables formal validation of the produced marked-up documents. While it is, of course, possible to develop an idiosyncratic schema that covers exactly the needs of a particular edition or project, this is a non-trivial process, esp. with materials as complex as are text-critical editions. It is thus much easier – as well as leading to better results – to employ a standard schema for a particular document type, as long as one is available.

The Text Encoding Initiative Guidelines [1] are specification primarily meant for scholarly encoding of texts. The TEI is an open de-facto standard, with a substantive history and large user community. It covers a wide variety of text and annotation types: in particular, it offers tagsets for prose, verse, drama, and dictionaries; for textual criticism, transcription of primary sources, linguistic analysis, and more. It also offers a header definition, which allows the inclusion of detailed metadata, as well as extension and modification mechanisms.

Given its generality, the TEI does not define one single schema to cover all types of materials and types of annotation. Rather, the current version (TEI P4) offers a number of modules, which can be combined (and further extended) to arrive at a particular schema realised as an XML Document Type Definition (DTD). For our project we chose the following modules:

- TEI.prose, the base module for encoding prose – it contains elements for the TEI header, giving detailed meta-data on the document (such as file, source, encoding and revision descriptions), as well as standard elements for document structuring (division, paragraph, table, note, …) and sub-paragraph annotation (emphasis, highlight, …);
- TEI.transcr, an additional module for the transcription of primary sources, in particular manuscripts, which includes elements for correction and emendation, recording the different hands in the text, etc.
- TEI.linking, an additional module that enables intra- and inter-document linking and contains elements and attributes to tie together the different transcriptions of the material and link the material with external resources;
- TEI.figures, an additional module to encode figures and other graphical material, used for encoding the facsimile, i.e. links to the graphic files containing the facsimile in various sizes and resolutions;
- TEI.extensions, a user specified module, that implements user extensions to the standard TEI – we used it to define some extra elements and to enumerate the attribute values for e.g. placement information on notes.

However, even after the TEI is parameterised for a certain project (i.e. we choose the required modules and extensions, arriving at an XML DTD), there is still considerable leeway in the choice of particular elements to use. Such a TEI DTD will contain, at least in the current version P4, also a large number of elements and attributes not required for the material. Such an over permissive DTD can, in the main, be useful for validation and interchange, but does, however, have a negative impact on authoring, making it difficult to use a DTD-aware menu-driven XML editor on the materials. This is why we – after defining the TEI parameterisation – also produced a strict (minimal) DTD, which specialises the TEI one. This DTD was used in the developmental cycle – in the final, public version of the materials we then revert to the “official”, TEI compliant DTD.

In Figure 3 we see an example of some material encoded according to our DTD. As can be seen, the elements are heavily indexed, with each division, page and line having its ID as well as the reference to the corresponding ID from another transcription

```
<div id="sl1d" corresp="sl1k" n="1" type="dipl">
  <head>Diplomatični prepis</head>
  <page id="sl1d-f.1" corresp="sl1f.1" n="1">
    <line id="sl1d.1" corresp="sl1k.1" n="1" rend="right">1825. XIII</line>
    <line id="sl1d.2" corresp="sl1k.2" n="2" rend="center">Na 16 nedelo po Binkuſhtih.</line>
    <line id="sl1d.3" corresp="sl1k.3" n="3" rend="center">K'kerſhanſkimu govorjenju</line>
    <line id="sl1d.4" corresp="sl1k.4" n="4" rend="center">nagovor.</line>
    <line id="sl1d.5" corresp="sl1k.5" n="5">Takrat bode tebi zheſt, kader tijiſti, kateri je tebe
po&#301F;</line>
    <line id="sl1d.6" corresp="sl1k.6" n="6">vabil, tebi porezhe: Prijatel, pomekni ſe gori.
      <note place="right">Luk. 14.</note></line>
    <line id="sl1d.7" corresp="sl1k.7" n="7" rend="center">Vvod.</line>
    <line id="sl1d.8" corresp="sl1k.8" n="8"><note place="left">1.</note> Tolko ſtánov je na le temu
zhaſnimu ſvetu, pa <emph>vſih</emph> le</line>
    <line id="sl1d.9" corresp="sl1k.9" n="9">eden pokliz: doſezhi nebeshko kraleſtvo.</line>
    <line id="sl1d.10" corresp="sl1k.10" n="10">Od ſvetliga zeſarja na ſedeshu ſlatim do
      <del hand="AMS">berazha</del> ſroma&#301F;</line>
    <line id="sl1d.11" corresp="sl1k.11" n="11">ka per palzi beraſhki je vſaki ſtan Bog is'volil</line>
...
```

Figure 3: A facsimile transcription in the canonical TEI/XML format – compare with Figure 2.

4.2 XSLT filters

As mentioned, there are two stages where conversion scripts are used on the data, i.e. the up-conversion from the digital source into TEI XML, and the down-conversion from the TEI XML into HTML.

For the up-conversion the data are first converted into a "sane" format, e.g. from Word into XML via any of a number of RTF to XML converters, such as OpenOffice or UpCast. Next, for each edition, dedicated transforms were written, which take the presentation-oriented source XML and convert it in a pipeline into the target TEI encoding. These filters were written mainly in XSLT, the XML transformation language, also a recommendation of W3C and hence a standardised specification which is supported by various tools, e.g. IE Explorer. However, while XSLT is ideal for encoding XML structure conversions, it is less suitable for cases where certain string patterns should give rise to XML structures. For such cases filters were written in the Perl programming language.

Down-conversion into HTML was also realised in XSLT, for each material separately, although here the various down-translations share substantial portions of their code. In addition to graphically realizing various elements of the TEI source (such as changing TEI <del> elements into HTML <s>, i.e. strikethrough), the down-conversion also makes the table of contents, and, crucially, produces a side-by-side view of the facsimile and text, and parallel views of the different transcriptions.

An important point for the scholar who wishes to know more about the (digital) edition is also the HTML rendering of its TEI header. This XSLT template expands the header tags into their localised string descriptions (e.g. <respStmt> to "Responsibility statement" or "Izjava o odgovornosti") and furthermore links each tag to its definition in the TEI Guidelines. As the TEI header also contains a list of the tags used in the body of the document, this means that all elements used in the material have directly available documentation.

While the viewing of the materials could be made much more flexible, the current set-up does have the advantage that it can be used off-line (the complete TEI Guidelines are also mirrored with each book) without the need for any software, save a Web browser.

4.3 Character encoding

A special issue is the complex (historical, phonetic) characters needed for the presentation of the materials. Many such characters are in fact supported by Unicode, but not all. For these we use the Private Use Area of Unicode, and a special publicly available ZRCola character set and font [5].

5. Results

The main result of our endeavours is the Web library of critical editions of Slovene literature at http://nl.ijs.si/e-zrc/. It currently contains its first three critical e-editions, in particular the Three Sermons on Language by Anton Martin Slomšek (1800–1862), including facsimile, diplomatic and critical transcriptions and notes [6]; a part of Sigismund Zois' (1747–1819) correspondence, including facsimile, the diplomatic transcription, and translation into Slovene (the correspondence is in German), as well as notes and a hyperlinked glossary of person names appearing in the letters; and a collection of poems by Alojz Gradnik (1882–1967), which contain transcriptions of 15 different variants (various printings as well as author's corrections) of the collection. Currently we are working on a number of other editions, where the most important (and complex) one is the "Freising Manuscripts" (972-1039), three religious texts, which are the oldest written Slovenian texts and the oldest Slavic texts written with Latin alphabet. Due to their importance, the critical edition encompasses an enormous apparatus: it includes facsimile, diplomatic, critical and phonetic transcriptions; translations into Latin, Old Church Slavonic and five modern European languages; and a dictionary covering the critical transcription, where each entity of dictionary contains the phonetic representation, grammatical information, translations, concordances from the critical transcription, and more. Additionally, the printed edition contains introductions, notes, and a bibliography.

A distinguishing feature of our library is the free availably of the materials (apart from the Gradnik poems, where this is not possible due to copyright restrictions) – not only are the editions available to everybody for browsing in their HTML form, but also for downloading as the TEI/XML source, with the accompanying facsimile graphical files. Such free access is possible as the original texts are usually over a hundred years old, while the authors and editors of the transcriptions and markup have agreed to make them freely available.

6. Benefits

The most evident benefit of our e-library is, of course, its accessibility – rather than having to buy or borrow a book, anyone with an internet connection can at any time peruse the materials. Furthermore, these texts, when presented in multimedia, reveal their historical dimension clearer and in a more attractive way than they would in printed form. An important advantage of the e-editions is the parallel presentation of the various transcriptions, which juxtaposes the original text with its more understandable forms. We are also working on further multimedia extensions, in particular the inclusion of audio streams with recorded recitation for each passage of the Freising Manuscripts. In this way, the pupil or student can get a clear and at the same time many-sided and dynamic idea of the historical development of language and national culture. These, in our opinion, are the most visible benefits of a complex, text-critical e-edition of an old text for pedagogical purposes on different levels of education, as well as for further research.

7. Conclusions

The communicative power of old texts and historical documents can be brought to their full expression in a synergism of their visual, full-textual and audio presentations. This goal, which is of considerable importance for such spheres as education, museology, archives, human studies etc., can be achieved by digitization and encoding of materials, strictly applying open encoding standards and making the resulting materials available on the Web. The paper presented the methodology and technology to achieve this aim by: 1) up-conversion to TEI/XML/Unicode via a collaborative and cyclic process of step-wise refinement, largely implemented by means of XSLT transforms, 2) down conversion into a user-friendly HTML mounted on a publicly accessible URL. This methodology has been tested on three completed e-editions, while several others are currently in the process of production.

We have aimed to make the editions maximally useful also in terms of availability, which is reflected in the extremely liberal access to the materials, with unrestricted copying of the source XML format. This approach is similar to the well-known Open Source software, where the materials can also be used for producing commercial, in our case printed or CD-ROM editions. As the probability of substantial revenue is virtually nil, and as our work has already been financed by government grants, such a maximally open licence seems to us to best further one of the important goals of the project, i.e. to achieve maximal availability of the editions.

As mentioned, further work includes the addition of audio streams, where we initially plan to connect manually segmented sound files directly to various milestone elements in the text, (e.g. to phrase boundaries marked in the phonetic transcription of the Freising Monuments) making the result immediately usable in the HTML format. Later, in the case of more complex multimedia content, we plan to use the SMIL standard [7]. In addition to sound, we will make the first attempt at including a video recording into our edition of the baroque-era Passion play from Škofja Loka. Another direction we would like to pursue in our further work is the addition of mark-up for linguistic structure to our texts [8,9]. This would enable the inclusion of the texts into a web-based concordancing engine (as already implemented at http://nl2.ijs.si/ for other corpora), as well as the extraction of parallel lexica from the transcriptions. All these analytical tools would contribute to expose the inner complexity, research potentials, and historical value of the texts published in our digital editions.

References

[1] Sperberg-McQueen, C. M., Burnard, L. (eds.) (2002). Text Encoding Initiative: Guidelines for Electronic Text Encoding and Interchange, TEI P4, the XML-compatible edition. TEI Consortium. http://www.tei-c.org/P4X/

[2] Hockey, S. (2000). Electronic Texts in the Humanities. Principles and Practice. Oxford University Press.

[3] Wymer, K. (2005). Why Universal Accessibility Should Matter to the Digital Medievalist. *Digital Medievalist* 1(1). http://www.digitalmedievalist.org/

[4] Schreibman, S., Kumar, A., McDonald, J. (2003) The Versioning Machine. *Literary and Linguistic Computing* 18(1), 101-107. http://mith2.umd.edu/products/ver-mach/

[5] Weiss, P. (2004). Vnašalni sistem ZRCola. (The text input system ZRCola) In *Language Technologies: Proceedings B of the 7th Intl. Conf Information Society, IS 2004*. Ljubljana: Jožef Stefan Institute, p.124. http://zrcola.zrc-sazu.si/

[6] Erjavec, T., Ogrin, M., Faganel, J. (2005). E-Slomšek: A TEI Encoding of a Critical Edition of 19th Century Slovenian Rhetoric Prose. *Review of the National Center for Digitization.* 6(4), 31–41.

[7] W3C. The Synchronized Multimedia Integration Language (SMIL). http://www.w3.org/AudioVideo/

[8] Erjavec, T. (2002). The IJS-ELAN Slovene-English Parallel Corpus. *International Journal of Corpus Linguistics*. 7(1), 1-20.

[9] Erjavec, T., Džeroski, S. (2004). Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence* 18(1), 17-40.