Izvorni znanstveni članak

Towards a Digital Edition of the Slovenian Biographical Lexicon

Petra Vide Ogrin
Slovenian Academy of Sciences and Arts, Library
Novi trg 3, Ljubljana, Slovenia
petra.vide@zrc-sazu.si

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Summary

The paper presents work-in-progress in the project of the digitization of the Slovenian Biographical Lexicon (SBL). SBL summarizes the lives and work of notable figures from Slovenia's history and is an important reference work for research in the Slovenian humanities and social sciences as well as in the history of the natural sciences. SBL uses the bio-bibliographical method and gives a synthetic assessment of their work and significance, based on the extensive use of primary sources, thus offering support to other encyclopaedic and reference editions. Today especially the first few volumes are almost impossible to obtain while, at the same time, requests for copies of SBL are frequent not only in Slovenia but also from abroad. In order to widen the availability of SBL, the Slovenian Academy of Sciences and Arts (SASA) and the Scientific Research Centre of the SASA have started a project aimed at digitising the SBL and making it freely available on-line. The methodology hinges on the use of open standards and software, in particular the XML-based Text Encoding Initiative Guidelines, which provide a wide, formally specified and well-documented annotation vocabulary for texts and analytical mechanisms. The paper presents the general methodology and steps of the SBL digitization work, concentrating on the TEI elements to be used in the digital edition, in particular: structural mark-up, identification of names and other biographical data, expansion of abbreviations, and cross-linking. The paper gives motivation for our choices and exemplifications from the SBL, and concludes with plans for further work.

Keywords: digital library, encyclopaedias, TEI-XML mark-up

Introduction

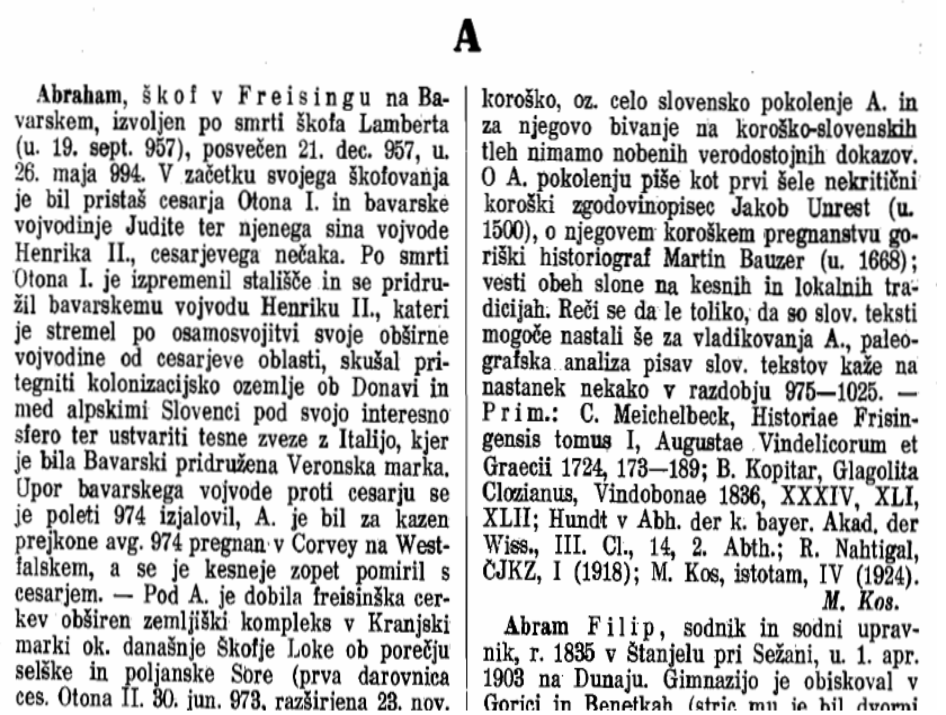
Publication of the Slovenian Bibliographical Lexicon (SBL, 1925-1991) started in 1925 with the appearance of the first of 15 volumes. The introduction to the first volume explains the motivation for such a publication: to give an accurate picture of Slovenia's cultural life, from its beginnings up to the contemporary time by including everybody who participated in the cultural development, is of Slovenian nationality or born on Slovenian ground, and was active in the homeland or abroad, as well as persons of foreign origin who with their work among the Slovenians influenced the Slovenian cultural life. SBL aims to cover not only a person's biography but also the important literature depicting the person's life and work, information where to find their unpublished works or photographs, in short, the SBL aims to be a reliable signpost in orientation and a helpful tool in scientific research.

The first editorial board proposed a list of persons to be included in SBL, which comprised 2,335 names who mostly covered the areas of humanities and social sciences, although the awareness of the necessity to include also figures from other areas was increasingly present. In its rather long history the publication of SBL saw various changes regarding the list of persons to be included. The new list proposed after the WW2 was supposed to "reflect the shifts in the society and a broader concept of national culture that should consider the increasing development of natural sciences, modern technologies and their applications, as part of the spiritual superstructure" (SBL, vol. 15, 1991). In the end, and despite the fact that a few dozens of the names from the original list were eliminated as their significance was examined anew, this led to the enlargement of the original number. So with the last volume having been published in 1991 SBL comprises 5,031 biographical entries and, since some entries are family names, SBL covers more than 5,100 persons, from the beginnings of the Slovenian culture up to the time after WW2.

The primary aim of the SBL was to be informative as well as exhaustive, so a balance had to be found between the length of an article and substantial information. The data in articles are checked against relevant historical materials, e.g. biographical and other dates are always compared to dates in registers and other primary documents, literary citations with originals, etc. SBL also includes an index of all person names that appear in the articles and a list of abbreviations.

In order to widen the availability of SBL, the Slovenian Academy of Sciences and Arts (SASA) and the Scientific Research Centre of the SASA have started a project aimed at digitising the SBL and making it freely available on-line. This paper explains the first steps in this direction, concentrating on the underlying methodology, and the information we plan to explicitly mark in the digital edition. The rest of this paper is structured as follows: Section 2 gives account of the TEI P5 Guidelines and outlines the changes they bring in markup practice, Section 3 gives account of the methodology used in the digitization process of

SBL, and Section 4 sketches the possibilities of the future work on the project of the digital SBL.



A

Abraham, škof v Freisingu na Bavarskem, izvoljen po smrti škofa Lamberta (u. 19. sept. 957), posvečen 21. dec. 957, u. 26. maja 994. V začetku svojega škofovanja je bil pristaš cesarja Otona I. in bavarske vojvodinje Judite ter njenega sina vojvode Henrika II., cesarjevega nečaka. Po smrti Otona I. je izpremenil stališče in se pridružil bavarskemu vojvodu Henriku II., kateri je stremel po osamosvojitvi svoje obširne vojvodine od cesarjeve oblasti, skušal pritegniti kolonizacijsko ozemlje ob Donavi in med alpskimi Slovenci pod svojo interesno sfero ter ustvariti tesne zveze z Italijo, kjer je bila Bavarski pridružena Veronska marka. Upor bavarskega vojvode proti cesarju se je poleti 974 izjalovil, A. je bil za kazen prejkone avg. 974 pregnan v Corvey na Westfalskem, a se je kesneje zopet pomiril s cesarjem. — Pod A. je dobila freisinška cerkev obširen zemljiški kompleks v Kranjski marki ok. današnje Škofje Loke ob porečju selške in poljanske Sore (prva darovnica ces. Otona II. 30. jun. 973, razširjena 23. nov.

koroško, oz. celo slovensko pokolenje A. in za njegovo bivanje na koroško-slovenskih tleh nimamo nobenih verodostojnih dokazov. O A. pokolenju piše kot prvi šele nekritični koroški zgodovinopisec Jakob Unrest (u. 1500), o njegovem koroškem pregnanstvu goriški historiograf Martin Bauzer (u. 1668); vesti obeh slone na kesnih in lokalnih tradicijah. Reči se da le toliko, da so slov. teksti mogoče nastali še za vladikovanja A., paleografska analiza pisav slov. tekstov kaže na nastanek nekako v razdobju 975—1025. — Prim.: C. Meichelbeck, Historiae Frisingensis tomus I, Augustae Vindelicorum et Graecii 1724, 173—189; B. Kopitar, Glagolita Clozianus, Vindobonae 1836, XXXIV, XLI, XLII; Hundt v Abh. der k. bayer. Akad. der Wiss., III. Cl., 14, 2. Abth.; R. Nahtigal, ČJKZ, I (1918); M. Kos, istotam, IV (1924).

M. Kos.

Abram Filip, sodnik in sodni upravnik, r. 1835 v Štanjelu pri Sežani, u. 1. apr. 1903 na Dunaju. Gimnazijo je obiskoval v Gorici in Benetkah (stric mu je bil dvorni

Figure 1: An SBL page excerpt

Text Encoding Initiative P5

In 2007, it has been exactly twenty years since the beginning of Text Encoding Initiative (TEI), whose main remit was to produce recommendations or guidelines for the creation and processing of electronic texts for better interchange and integration of scholarly textual data in all languages and from all periods (Burnard, 1988). The first public version of the TEI Guidelines, so called TEI P3 was published in 1994 (Sperberg-McQueen, Burnard, 1994) and was based on the ISO standard SGML. TEI P4, published in 2002 (Sperberg-McQueen, Burnard, 2002) converted the underlying annotation scheme to XML, although it still maintained backward compatibility with P3, and hence SGML. TEI P5 Guidelines, which are to be finalized in 2007, do not maintain compatibility with P4, and are thus free to explore new solutions to text encoding.

For an XML document to be valid (not merely well-formed), its structure must be checked against a XML schema. For a valid TEI document, this schema must be a conformant TEI schema, which is a project-specific combination of TEI modules. Each module contains a set of related elements, typically for use to describe a particular text type or analytical approach.. Typically in building a

TEI schema we decide for a combination of different modules, according to our needs. The TEI infrastructure module, however, is a required component of any TEI schema. It provides declarations for all datatypes, and for the attribute classes, model classes, and macros used by other modules in the TEI scheme.[1] To define a schema in P5, i.e. the required combination of TEI modules and their customizations, the so called ODD language is used.[2].
Apart from the extension and generalization of a modular system one of the main goals of TEI P5 Guidelines is interoperability with current standards, in particular those of ISO and W3C. TEI encoders, for example, are mandated to use Unicode characters. Similarly, the old global attribute lang has been replaced by the W3C-defined attribute xml:lang, which takes as its value the ISO language code. P5 also brings some new elements which are especially relevant for the digital edition of SBL, in particular a structured description of real world entities such as persons and places independent of textual references to them. In this way, "the scope of TEI encoding scheme expands beyond the simple representation of textual structure to include the representation of the knowledge inferred from or implicit in those textual structures." (Burnard, 2007).

Methodology

The methodology of encoding SBL hinges on the use of open standards and software, in particular the adoption of the TEI P5 Guidelines. 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 XHTML Web presentation language, along with the implementation of digital library open source software. This software should be able to process users' search demands regarding full-text search and allow for complex search and display tasks combining different criteria. In the following sections we detail these aspects of the technology used in preparing the electronic edition of SBL**.**

Preparation of the text

An SBL article typically starts with a name, which gives the variant of the name, used by the person him/herself towards the end of his life or that was generally used, whereas other variants, if they exist, are put in brackets. What follows is a chronological summary of a person's life and activity: information about birth, death, residence, occupation etc. An article is mostly concluded with a brief bibliography that depicts the person's life and work or with locations of other materials relevant to the person's life, e.g. photographs. An article may consist of one or more paragraphs, depending on the exhaustiveness of the

[1] http://www.tei-c.org/release/doc/tei-p5-doc/html/ST.html#STIN

[2] One Document Does it all; see http://www.mulberrytech.com/Extreme/Proceedings/html/2004/Burnard01/EML2004Burnard01.html

article. As is characteristic for encyclopaedic texts, the SBL articles are written in a dense language, containing many abbreviations. Figure 1 shows the opening page of SBL with part of its first article.

In the process of encoding we have to pay special attention to two types of abbreviations: those that are contained in already existing lists, and those which represent the initial letter of a name entry. There are three existing lists of abbreviations with their expansions: a list of bibliographical abbreviations, a list of abbreviated names of SBL contributors and a list of general abbreviations. The lists changed with the publication of each volume in accordance with its content, so, for the purpose of our project, all lists from every volume are combined.

The text is first prepared in a text-editor with the prospect of further conversion into XML. The styles of the original text are retained in order to preserve some basic encoding in the up-conversion, such as <div> for articles and <p> for paragraphs. The text-editor we use is OpenOffice,[3] an open source office tool, which is able to import and export arbitrary XML, if provided with XSLT stylesheets to describe the transformation. TEI provides a simple set of stylesheets and templates to let TEI-XML and OpenOffice work together, in the TEI OO[4] package. So, after correcting the mistakes that still remain after OCR, text is converted into the basic TEI-XML format. At this point we consider which information we want to encode or which "intelligence must be embedded in the text in such a way that the program can derive information from it", (Hockey, 2000, p. 24) The structure of a TEI-XML document takes into account the encyclopaedic nature of the text and with the prospect of what information we want the user of our final database to be able to retrieve and in what way we want it to be presented.

TEI mark-up

The structure of SBL TEI-XML document follows the latest version of TEI P5 Guidelines[5] and especially the module on biographical and prosopographical[6] data. The module defines some special purpose elements which can be used to encode biographical, historical, and prosopographical data. The guidelines propose also a possibility of a more structured entry, for which the information contained in the text must be first extracted and then encoded separately using specific elements. The principal idea is that each article is treated individually

[3] http://www.openoffice.org/

[4] http://www.tei-c.org/Software/teioo/

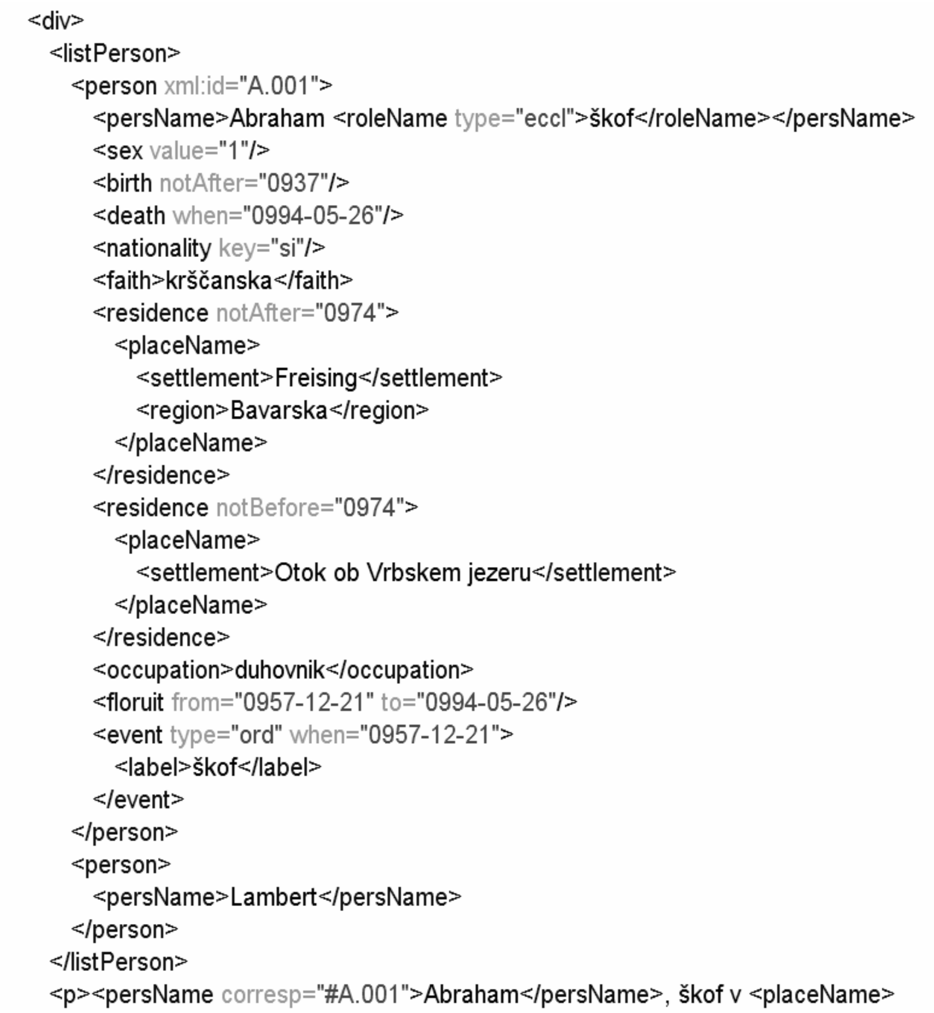
[5] http://www.tei-c.org/release/doc/tei-p5-doc/html/ND.html#NDPERS

[6] Prosopography: a study that identifies and draws relationships between various characters or people within a specific historical, social, or literary context (Gk. *prosopon* = person, face; *graphein* = (to) draw, write)

as a unit. Essential information about the subject of an article contained in the text is extracted and encoded in a separate structural block at the beginning of the article, in which the <person> element contains detailed further information.
Following TEI P5 Guidelines, information about people may comprise a series of statements or assertions that mainly fall into three categories: traits, states and/or events. Traits do not, by and large, change over time, and are encoded in elements such as <sex> or <trait>; states hold true only at a specific time and are encoded in elements such as <occupation> or <floruit>; events or incidents lead to a change of state or, less frequently, trait and are encoded as <event>, <birth> etc. The choice of TEI elements in SBL articles depends on the information contained in a particular article, so the number and the types of elements are to some extent different for each individual article. Nonetheless, we decided for a selection of typical elements that will be important for IR later on by the users, e.g. marriage, ordination, exile, further education, occupation, residence, active period, e.g. as a writer, doctor, judge, bishop etc.

Figure 2 shows the elements used for the first article of SBL. Within the <person> element there are elements indicating sex, date of birth and death (in this case date of birth is not certain), nationality, faith, residence, occupation, active period, occasion of ordination and the role of bishop. It should be noted that the values of attributes are ISO or W3C standard wherever possible (sex, dates etc.). The elements in the <listPerson> are to be (semi-)automatically extracted from the annotation in the article itself. So the basic principle underlying the encoding process of the SBL is first to encode the information in the article with the appropriate elements and then capture it within the <listPerson> element. More on the reason for such doubling of metadata in the section on information retrieval.

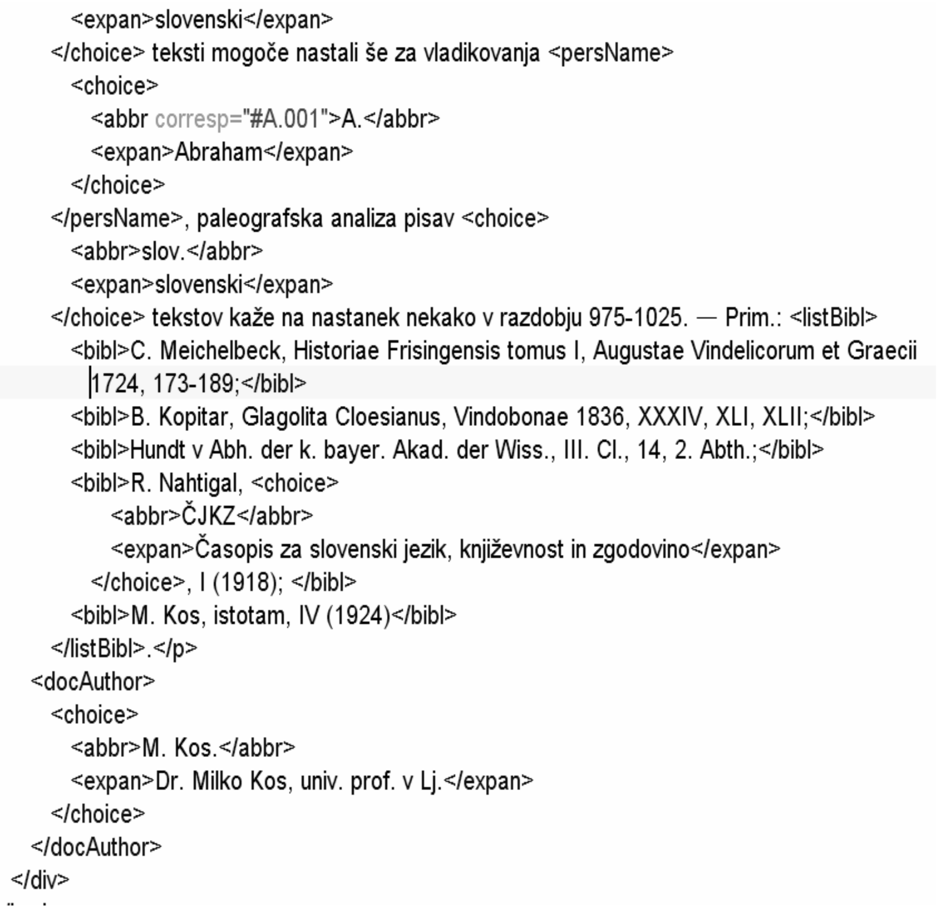
The elements also contain linking attributes that join elements conveying the same content or in some other way corresponding to each other. Cross-linking may connect a biographical entry with the occurrences of the name or name variants elsewhere in the entire SBL text and to the entry in the index of personal names. The elements of <occupation> and <floruit> are connected by virtue of adjecency and give information about which occupation a person had in a certain period or give indication of the field of activity in that period. Explicating such information will ideally allow a user to form search demands such as: when did so and so practice law or what kind of interests someone exhibited between 1850 and 1880.



```
<div>
  <listPerson>
    <person xml:id="A.001">
      <persName>Abraham <roleName type="eccl">škof</roleName></persName>
      <sex value="1"/>
      <birth notAfter="0937"/>
      <death when="0994-05-26"/>
      <nationality key="si"/>
      <faith>krščanska</faith>
      <residence notAfter="0974">
        <placeName>
          <settlement>Freising</settlement>
          <region>Bavarska</region>
        </placeName>
      </residence>
      <residence notBefore="0974">
        <placeName>
          <settlement>Otok ob Vrbskem jezeru</settlement>
        </placeName>
      </residence>
      <occupation>duhovnik</occupation>
      <floruit from="0957-12-21" to="0994-05-26"/>
      <event type="ord" when="0957-12-21">
        <label>škof</label>
      </event>
    </person>
    <person>
      <persName>Lambert</persName>
    </person>
  </listPerson>
  <p><persName corresp="#A.001">Abraham</persName>, škof v <placeName>
```

Figure 2: TEI-XML structure of extracted information in <listPerson

Special attention is paid to handling the problem of abbreviations. As mentioned above, there are three types of abbreviations and the comprehensive existing lists give the expanded form or explanation for each abbreviation. We wrote Perl programs to automatically tag the abbreviations in the text with suitable elements, as shown in Figure 3. A program is also used to identify and encode the initials of the subject of an article that occur further in the text of an article contained in <p> element(s). However, it should be noticed that the expansion of the abbreviations is, as yet, preliminary, as they would still need to be correctly inflected.



```
      <expan>slovenski</expan>
    </choice> teksti mogoče nastali še za vladikovanja <persName>
      <choice>
        <abbr corresp="#A.001">A.</abbr>
        <expan>Abraham</expan>
      </choice>
    </persName>, paleografska analiza pisav <choice>
      <abbr>slov.</abbr>
      <expan>slovenski</expan>
    </choice> tekstov kaže na nastanek nekako v razdobju 975-1025. — Prim.: <listBibl>
      <bibl>C. Meichelbeck, Historiae Frisingensis tomus I, Augustae Vindelicorum et Graecii
        1724, 173-189;</bibl>
      <bibl>B. Kopitar, Glagolita Cloesianus, Vindobonae 1836, XXXIV, XLI, XLII;</bibl>
      <bibl>Hundt v Abh. der k. bayer. Akad. der Wiss., III. Cl., 14, 2. Abth.;</bibl>
      <bibl>R. Nahtigal, <choice>
          <abbr>ČJKZ</abbr>
          <expan>Časopis za slovenski jezik, književnost in zgodovino</expan>
        </choice>, I (1918); </bibl>
      <bibl>M. Kos, istotam, IV (1924)</bibl>
    </listBibl>.</p>
  <docAuthor>
    <choice>
      <abbr>M. Kos.</abbr>
      <expan>Dr. Milko Kos, univ. prof. v Lj.</expan>
    </choice>
  </docAuthor>
</div>
```

Figure 3: Annotation of abbreviations

Retrieval of biographical information

The idea of placing essential information in a special structural block within the <listPerson> element preceding the text of an entry organizes the information and so makes it accessible in a well defined way. Information captured in this sort of unified metadata structure is analogous to a TEI Header or a bibliographic record, where data are placed into a variety of predefined fields and then made searchable by those fields. Such a system addresses demanding users, having specifically defined research needs. In the case of SBL a user will be able to find for example female writers, who lived and worked in the period between 1830-1860 in Ljubljana, or priests, who were also philosophers and born in Maribor etc.

Future plans

The project of the digital edition of SBL is still at a very early stage. Our future work will concentrate on various aspects of the SBL digitization. First we will work on the implementation of a system that will allow for the kind of advanced searching described above. One possibility considered is PhiloLogic[7], a full-text search, analysis, and retrieval tool developed by the ARTFL Project[8] and the Digital Library Development Center at the University of Chicago. The system supports boolean and proximity searches and provides a number of reporting functions, including KWICs, and achieves quick display of results by pre-indexing each word and storing byte offsets in a flat file. PhiloLogic handles document structure, currently extending down to seven levels of depth, abstracting a structure by assigning numbers to each nested object-level of each document. It uses an abstract representation of document structure, shredding the XML into sets of related database tables. This means that PhiloLogic can process a TEI-XML encoding scheme by extracting structural information from available text tagging. Storing object-level data in SQL tables, PhiloLogic can search document structure and refine word searching by using the shredded XML. (Cooney et al., 2007)

Apart from working on the implementation and adaptation of the retrieval system as a whole, we will concentrate on semi-automatic tagging of the text. Dedicated filters written in the Perl programming language and XSLT will be able to produce further mark-up in the up-conversion from OpenOffice into TEI. XSLT, which is the XML transformation language, also a recommendation of W3C and hence a standardized specification, is ideal for defining XML structure conversions. It is less suitable though for cases where certain string patterns should give rise to XML structures and in such cases filters are written in Perl (Erjavec, Ogrin, 2005).

Semi-automatic annotation also means exploring the field of language technologies, especially in the extraction of various biographical data with specific consideration of Slovenian linguistic particularities, e.g. inflected forms that need to be normalised, or, as in the case of abbreviations, lemmas that need to be inflected. Regarding the character of data the emphasis will first be probably on the development of a Named Entity Recognition (NER) tool, (Jackson, Moulinier, 2002; Bekavac, 2002) for Slovenian, which will undoubtedly speed up the encoding process.

[7] http://philologic.uchicago.edu/

[8] http://humanities.uchicago.edu/orgs/ARTFL/

Conclusions

The motivation for digitization of the SBL is for the research and other community to be able to access the valuable reference information captured in the lexicon. The digital medium offers searching and retrieval possibilities unthinkable to a classical reader of printed text, especially regarding a text of such encyclopaedic nature as is SBL. Thus, the main idea is to produce a digital edition of SBL with an implementation of a satisfactory retrieval system that will be able to process intelligent searching tasks. To this purpose we have decided for a more complex TEI-XML structure of a biographical entry. That often means (semi-)manual extraction of data that are not explicitly expressed in the text itself. Adding such metadata may mean, in some way and to some extent, creating “new” data, or making implicit information explicit, which represents further editorial interventions. We expect future development of linguistic tools for data extraction to prove helpful in being able to do automatic data extraction, which will speed up the process of encoding.

We believe the implementation of a suitable open source digital library software together with the TEI P5 encoding as outlined in this paper will answer the needs of a demanding user.

References

Bekavac, Božo. Strojno obilježavanje hrvatskih tekstova – stanje i perspektive // *Suvremena lingvistika.* 53-54 (2002), 173-182. http://nl.ijs.si/et/tmp/matija/BB_disertacija_verT9.pdf (2007-09-14)

Burnard, Lou. Encoding standards for the electronic edition // Znanstvene izdaje in elektronski medij: razprave / Ogrin, M. (ed.). Ljubljana: Založba ZRC, ZRC SAZU, 2005, 25-42

Burnard, Lou. Report of Workshop on Text Encoding Guidelines // *Literary & Linguistic Computing*. 3 (1988), 131-133

Cooney, Charles M. et al.. Extending PhiloLogic. April 2007.

http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=175 (2007-09-09)

Erjavec, Tomaž; Ogrin, Matija. Digital Critical Editions of Slovenian Literature: an Application of Collaborative Work Using Open Standards. // *From Author to Reader: Challenges for the Digital Content Chain: proceedings of the 9th ICCC International Conference on Electronic* Publishing, Arenberg Castle / Dobreva, M.; Engelen, J. (eds.). Leuven: Peeters, 2005, 151-156

Hockey, Susan. Electronic Texts in the Humanities: Principles and Practice. Oxford: University Press, 2000

Jackson, Peter; Moulinier, Isabelle. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. Amsterdam: John Benjamins, 2002, 180-185

Cankar, Izidor et al. (eds). Slovenski biografski leksikon. Ljubljana: Slovenska akademija znanosti in umetnosti, 1925-1991

Sperberg-McQueen, C. Michael; Burnard, Lou (eds.). TEI P4: Guidelines for Electronic Text Encoding and Interchange (TEI P4). Text Encoding Initiative Consortium. XML Version. Oxford, 2002. http://www.tei-c.org/Guidelines2/index.xml.ID=P4 (2007-09-09)