Digital Edition of Leo Tolstoy Works:
Contributing to Advances
in Russian Literary Scholarship

Anastasia A. Bonch-Osmolovskaya*

National Research University "Higher School of Economics"
20 Myasnitskaya Str., Moscow, 101000, Russia

The paper discusses the practices of digital editions in the context of digital heritage preservation. A brief overview of the state-of-art in Russia is proposed. It is stated that an up-to-date digital edition project should meet a number of requirements which concern conditions for long-term preservation, data accessibility and sustainability of formats. The paper sets a perspective to discuss a new project on Leo Tolstoy's literary heritage ("Tolstoy Digital"). Based on 90-volume complete edition of Tolstoy's work, the project, nevertheless, is positioned not as a digital reproduction of a printed edition but as a freestanding resource open for absorbing data from other sources.

1. Digital heritage preservation

1.1. Recent trends
in digital heritage preservation

Digital editions constitute one of the pivotal domains of digital humanities. They started in the early 90-s with establishing of virtual archives, opening access to rare printed editions or publishing primary texts. The practice of digital editions has currently evolved to a broad field that is nowadays more concerned with the issues of infrastructure and theory. One of the major questions under discussion is the problem of survival and longevity of digital documents. As pointed out in (Lee et al 2002), «Digital technologies enable information to be created, manipulated, disseminated, located, and stored with increasing ease. Ensuring long-term access to the digitally stored information poses a significant challenge, and is increasingly recognized as an important part of digital data management». A tremendous number of early projects developed in the 90-s have been lost due to technical progress that resulted in the change of formats and software (Earhart 2015). Therefore, one of the key challenges

*   Corresponding author E-mail address: abonch@gmail.com

nowadays is to overcome the dependency of digital objects on technological change.

At the end of 2014, the European Commission introduced a roadmap for preservation of digital cultural objects. The roadmap was developed as a part of the project supported by the European Commission in the frames of the Seventh Framework Programme for Research and Technological Development. The main provisions of the roadmap were reflected in the corresponding handbook (DCH-RP 2014). The document outlined the strategies for the preservation framework that would support long-term access to digital archives, compatibility of tools and software used for digital heritage development, security of storage, and the audit methods that validate trustworthiness of the repository.

The authors of the document describe a digital object consisting of three layers: a physical, logical and conceptual layer. The structure of each layer implies its own preservation method. *Bit preservation* is regarded as a number of basic actions, ensuring the integrity of code. *Logical preservation* ensures the quality of being able to reproduce the object and maintain it over time. It is closely connected with the requirements of format sustainability. *Semantic preservation* is responsible for the interpretability of the data in long-term perspective and transparency of annotation scheme and editor choices. The term 'preservation' means that there will always be an answer to questions of where one obtained the data from, how it can be interpreted, what the tags stand for, how this information may be linked or integrated within other sources. A desired level of "semantic security" can be only obtained through hard collaborative work of the community which develops and establishes appropriate standards for data and metadata annotation. A recognized way of text encoding is the standard developed by the Text Encoding Initiative, usually referred to as TEI .

1.2. Basic Principles of the Text Encoding Initiative

The TEI is a collectively developed and maintained standard used for the representation of texts in a digital form, as it is indicated on the front page of the TEI web site. The TEI is also the name of the consortium, engaged in the development of the standard, its documentation, dissemination and elaboration of accompanying software. The TEI started in 1987 when the representatives of more than 40 institutions and projects gathered to lay the foundations of a unique machine-readable standard for digital publications.

Now we have the fifth version of the TEI documentation (the TEI P5 guidelines), the first one dates back to as early as 1990. The TEI P5 guidelines are fully electronic, open access, and include 23 chapters that define coding of any possible type of textual, meta-textual and contextual information including variants, audio sources and the state of a material source. According to Lou Burnard, one of the TEI creators, "the TEI framework provides a useful way of thinking about the nature of text: it constitutes a kind of encyclopaedia of generally-agreed textual notions, using the vocabulary defined by the TEI in its Guidelines" (Burnard 2014:13).

The TEI language follows xml syntactic rules which require a declarative first line (namely, <TEI xmlns='http://www.tei-c.org/ns/1.0'>) and a structure set up by the use of start and end-tags that should be both present and should not intersect with other pairs of tags. The XML syntax makes TEI documents compatible with numerous software tools that are used to create, transform, analyze, or publish xml-files.

Every TEI-document consists of two parts: TEI-header marked by the tag <teiHeader> and

the text itself marked by the tag <text>. According to the guidelines, the TEI-header documents an encoded work, text's imprint or manuscript description, its title and author, its language, encoding, revisions, the names of digital editors. In other words, the TEI-header serves as a passport for a digital document.

The textual module may differ a lot depending on the genre of the text: a journal article, a novel, a poem, a drama, a letter exploit different tags which feature specific information relevant for the genre. For example, while coding a journal article you need to be most attentive to the reproduction of formatting, for example, distinguishing introductory paragraphs or lists from the main text. For a poetic text, the most important structural information is about line breaks. The TEI standard also provides a possibility for generous rhythmical mark-up. A dramatic text is created with a sort of an annotation used to connect speakers to their speeches, alternatively, they can be separate stage directions. This annotation gets a standard coding with TEI tags, such as <stage> for 'stage directions', or <sp> for 'speaker'. Correspondence is inseparable of its metadata, which is an addressee and a recipient, address, date and place of writing. This metadata is coded within the TEI frameworks as a separate block at the beginning of the document.

Named entities such as person, location, organization may be annotated within the text or may be kept as a separate list with a unique identification for each entrance. In this case a tag in the text provides only a reference link to the item from the list. The TEI has a specification for semantic annotation of the most important biographical events or social indicators, such as occupation, affiliation, residence, language knowledge etc. This module makes possible semantic coding (and, therefore, semantic search) of multifarious biographical information

that commentaries and biographical references generally provide.

Two other annotation modules to be discussed here are Critical Apparatus and Representation of the Primary Source. They are used to reflect the variations of the text and the layers of the document, corrections made by the author, mistakes and interpretations introduced by the editor. A digital copy, therefore reflects not only the primary document but also the author's and the reader's work of creation and comprehension. In other words, a digital edition constitutes a new object which is not equal to its material sources. Some valuable insights on this issue are given in (Robinson 2013).

"The text is the site of meaning which links the document and the work. The work can never have a fixed physical expression. It can only be apprehended (and ever only incompletely) in the text we construct from the document. The document without the text of the work we construct from it is mute, simply marks on a surface. Our construction of the text of the work, from one document, from a thousand documents, demands all our attention, all our knowledge, all we know of intention, agency, authority. There is no end to this knowing." (Robinson 2013:120).

This vision enhances considerably the function of a digital editor as the layers of reader's apprehension and knowledge coexist with the primary text in the digital document.

The TEI framework is used in numerous projects, some of them, such as "World of Dante" and "Decameron" are separate works, others are digital archives of famous people, such as Emily Dickinson, Van Gog, Henrik Ibsen, Jeremy Bentham and many others. Many projects are the collections of texts associated with a social group or cultural movement, see, for example "Victorian Women Writers' Project", "Wright American Fiction 1851-1875", "The Poetess Archive" and

many others. A detailed survey of different TEI projects is given in (Skorinkin 2016).

What makes the TEI framework so popular is not only the transparency and minimalism of its syntactic structure and exhaustive documentation, but also highly collaborative community which maintains the infrastructure: software, journal, events, SIGs, and online forums.

1.3. Digital preservation in Russia

The idea of data curation seems to be quite new for the Russian community although numerous initiatives in digital heritage have been recently launched in various parts of the country. The projects using the TEI standards or an adapted variant of the TEI can be about long-term sustainability of data. Five projects are definitely worth discussing here.

Manuscript

Manuscript is a digital archive of medieval Russian manuscripts supported by advanced search options for graphics and morphology (reference). According to Votintcev (2006), the project authors use an adjusted version of the TEI format for document annotations.

Anthology of the 18th century Russian literature

Th digital archive is a pilot platform for promoting online digital editing in Russia (Andreev et al 2009). Though small in size, the collection exploits the possibilities of TEI annotations. In addition to to some basic metadata, the documents are marked by important tags for various textual properties, such as themes, genre, rhythm, geo names, person names, mythological names etc.

The Archive of Bashkir Folklore seems to be the only resource, providing access to primary xml (TEI) files. This archive stores digitized field notes on folklore in Bashkir starting from 1950s.

In addition to concordance search, the archive provides frequency statistics on word forms and links every document to Google maps, showing places where the texts have been collected. (Orekhov et al 2012).

Two other projects use a light version of TEI (XML/TEI Lite).

Mandelstam's World is a virtual archive of Osip Mandelstam's works and it seems to be one of the soundest Russian digital humanities projects of Russian. It is aimed to pull together the poet's archives scattered all other the world – his manuscripts, variants of his poems, official and non-official documents on his life and work. The introductory page of the virtual archive gives us a detailed description of the the project's visionary goals and the technical tools used to achieve them. An extended project description can be found in (Nerler 2014). According to the paper, all the documents of the archive get a unified description in XML/TEI Lite framework, and this format will be used for their preservation. The web site presents the documents with facsimiles and metadata records, including some properties of the material source and indications of its physical location. Even despite a poor functionality of search, the project can still can be regarded as a great step forward in practices of data preservation in Russia.

The last project to be discussed here is the **Fundamental Electronic Library of Russian literature and Folklore (FEB)**. The FEB was launched in 2002 as an ambitious project supported by the Russian Academy of Science and the RF Ministry of Information Technologies. The basic unit of the FEB is a so-called Digital Scholarly Edition, which may be a work, an author or group of related texts, a critical apparatus or comments. The overall idea of the project is to represent the canon of Russian classical literature and literary scholarship of Russian writers. Though being indisputably remarkable and influential project

with highly elaborated search module, the FEB has some conceptual shortcomings that have been poignantly described in (Mjor 2009). As Mjor shows, some documents of the Soviet era cannot be interpreted properly without the historical context (for example, Literary Encyclopedia "which bears ideological imprint of Stalin's era" (Mjor 2009:94). Following its principle to present a precise digital version of the typography of the original, the FEB leaves no space either for "work", or for "text" in Robinson's terms (Robinson 2013).

Some requirements to a contemporary digital edition both of technical and conceptual nature are summed up in a short list below.

1) Exhaustive documentation should reflect the structure of the document corpus, its prospective and present condition, metadata properties and text annotations.

2) Document files should have a well defined and described format, and should be open for access and download if possible. If direct access is not possible, then their location and the names of scholars/curators responsible for their preservation should be specified.

3) Functionality of search should correspond to the declared document mark-up.

4) A digital edition is not equal to material sources; it should strive to represent knowledge backgrounds that enhance primary documents.

Tolstoy Digital project was launched in 2015 with an aim to present a digital edition of the 90 volume collection of Tolstoy's works. The conceptual vision of the project was inspired by the keen awareness of Russian culture's unjustified absence in the global digital landscape. The project does not only change this regrettable situation integrating a large part of Russian culture in the universal digital network, it also has a mission to promote up-to-date standards of digital editions. In this respect, Tolstoy's literary heritage seems to be an appropriate choice.

2. Digital edition of Tolstoy's complete works.

The 90 volume Tolstoy's complete works edition – so called "Jubilejnoe" (the Jubilee) – had been published for 30 years. The editorial work lasted from 1928 to 1958, the 90th volume with an index was published in 1964. The edition consists of three parts. Volumes 1 – 45 were for fiction and non-fiction works (previously non-published variants and drafts were also included), volumes 46 – 58 were for Tolstoy's diaries, volumes 59 – 89 included Tolstoy's correspondence.

It has been the most comprehensive publication of everything what Tolstoy has ever written, though, as the editors of a new, still unpublished, 100 volume edition point out, many gaps remain: the collection does not include numerous drafts, especially those that were hard to read, drafts of religious and philosophical treatises were published selectionally due to the twentieth-century censorship, let alone numerous mistakes made by copyists and type-setters. Leo Tolstoy's huge correspondence was not fully published, either (Opul'skaya 1997). Nevertheless, the Jubilee edition is a grandiose collection of Tolstoy's texts and reference materials and until recently it has been a bibliographical rarity.

In 2014, a unique crowdsourcing project "All Tolstoy in one click" was launched by the Tolstoy Museum in Moscow and one of Russian's top IT companies ABBYY, a leader in optical recognition. The 90 volume edition was digitized with the help of ABBYY OCR technology and then proofread by thousands of volunteers from 49 countries in two weeks. Now the works can be downloaded free of charge in popular e-books formats.

Furthermore, availability of xml files obtained as a result of the crowdsourcing project

enabled the idea to develop a full-pledged digital edition of Tolstoy's heritage.

The project's objectives may be stated in the following way:

1) to annotate all sorts of relevant data in Tolstoy's works, using the TEI-framework
2) to create a complete database of all named entities mentioned in the texts or commentaries
3) to link variants and drafts
4) to publish the results on the web providing an extensive search and visualization tools

A stumbling stone of the concept was to decide whether it should be a digital replication of the Jubilee volumes, though enhanced with semantic mark-up, or the complete edition of Tolstoy's works should be only used as a source for an independent digital resource. We preferred the second option. There are several arguments that favor such a decision. First of all, the Jubilee edition is a product of its time. This means that almost every volume contains an ideologically biased introduction that has only a slight historical relation to Tolstoy's work. Secondly, we have taken into account expert perceptions of editorial and typist imperfections. An independent approach is open to improvements and enrichments that may not be related to the Jubilee edition. Finally, though we understand that an idea of interlinking numerous texts (documents, correspondence and diaries of Tolstoy's contemporaries, biographies, bibliographies, literary scholarships etc.) into one global network resembles rather a distant dream than a clear prospect, we believe that conceptual foundations to develop this digital edition should be laid from the very beginning.

The work on the project can be divided into three stages: basic TEI implementation, creating a database with all named entities and their attributes, and embedment of textual interlinkage, including references and notes, textual variations, corrections, editorial interpretations, etc.

The first stage was about preliminary tasks. The decision to build a digital edition distinct from the 90 volume edition determined the TEI document specification: it was not equal to a volume but to a separate work. So, primary xml files of recognized and proofread volumes were cut into files with separate works or, alternatively, pasted together in case of grand oeuvres, such as "War and Peace". Diaries were not cut into separate entries. In this case, the TEI document is a writing-book, which is conceptually close to its original manuscript. Finally, letters were kept "in volumes", this meant that volumes 58-89 were not separated, each volume being considered as a TEI-document with a collection of letters.

After that, the xml-tags in the files were changed into corresponding TEI-tags and complemented with TEI-headers which contain necessary metadata (see Fig. 1).

Editors keptthe pre-revolutionary orthography used in the primary sources for some of the volumes. The problem resulting from this orthography is that it impedes morphological parsing as the word forms differ from those that could be found in the parsers dictionary. That is why, the next stage of data processing was bringing together all the words in old spelling new spelling (see Fig. 2).

The next stage of the project will be devoted to developing a biographical database. For this purpose, parsing of indexes at the end of each volume seems to be a nice starting solution. Most of valuable biographical information is contained in commentaries. We demonstrated elsewhere that a large portion of bio facts may be obtained automatically (Bonch-Osmolovskaya, Kolbasov 2015). Our plans are to link the database items to

```xml
▼<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xi="http://www.w3.org/2001/XInclude">
  ▼<teiHeader>
    ▼<fileDesc>
      ▼<titleStmt>
          <title>ДЕТСТВО</title>
          <author>Толстой Л.Н.</author>
        ▼<respStmt>
            <resp>подготовка TEI/XML</resp>
            <name>Евгений Можаев, Мария Картышева, Даниил Скоринкин</name>
          </respStmt>
        ▼<respStmt>
          ▼<name>
              Анастасия Бонч-Осмоловская, Фёкла Толстая, Борис Орехов
            </name>
            <resp>Идея, постановка задач, руководство</resp>
          </respStmt>
        </titleStmt>
      ▼<sourceDesc>
        ▼<biblStruct>
          ▼<analytic>
              <author>Толстой Л.Н.</author>
              <title level="a">ДЕТСТВО</title>
            </analytic>
          ▼<monogr>
            ▼<title level="m">
                Полное собрание сочинений. Серия первая "Произведения". Том 1
              </title>
            ▼<imprint>
                <pubPlace>Москва</pubPlace>
              ▼<publisher>
                  Государственное издательство "Художественная литература"
                </publisher>
                <date when="1935"/>
              </imprint>
            </monogr>
          ▼<series>
              <title level="s">Л.Н. Толстой. Полное собрание сочинений</title>
              <biblScope unit="vol">1</biblScope>
            </series>
          </biblStruct>
        </sourceDesc>
      ▼<publicationStmt>
        ▼<p>
            Проект
            <title>Толстой.Digital</title>
            разрабатывается сотрудниками и студентами
            <orgName>Высшей школы экономики</orgName>
            в сотрудничестве с
            <orgName>Государственным музеем Л.Н. Толстого</orgName>
            . Источник текстов –
            <bibl>90-томное собрание сочинений Л.Н.Толстого</bibl>
            . Разметка основана на стандарте
            <ref target="http://www.tei-c.org">TEI (Text Encoding Initiative)</ref>
            .
          </p>
        ▼<availability>
          ▼<p>
              Тексты и метатекстовая разметка доступны для свободного использования и распространения по лицензии Creative Commons Attribution Share-Alike (cc by-sa)
            </p>
          </availability>
```

Fig. 1. An example of TEI-header for "Childhood"

external web sites, such as Wikipedia, DBpedia and relevant Linked Open data archives.

Web access and accompanying services will be provided as soon as the first stage of the project is completed. The project's further development includes regular updates of the web portal, a part of Tolstoy's museum web site in Moscow (www.tolstoy.ru).

The data in the project is open and free to access and download, in line with Leo Tolstoy's ideas. The editors of the digital edition are always grateful for any contribution to the project.

```
              Государю Императору угодно освободить
▼<choice>
    <reg>помещичьих</reg>
    <orig>помѣщичьихъ</orig>
</choice>
▼<choice>
    <reg>крестьян</reg>
    <orig>крестьянъ</orig>
</choice>
  . Совершенно справедливо его
▼<choice>
    <reg>убедили</reg>
    <orig>убѣдили</orig>
</choice>
▼<choice>
    <reg>в</reg>
    <orig>въ</orig>
</choice>
▼<choice>
    <reg>том</reg>
    <orig>томъ</orig>
</choice>
  , что
▼<choice>
    <reg>крестьян</reg>
    <orig>крестьянъ</orig>
</choice>
  нельзя освободить иначе,
▼<choice>
    <reg>как</reg>
    <orig>какъ</orig>
</choice>
▼<choice>
    <reg>с</reg>
    <orig>съ</orig>
</choice>
  землей, на которой они
▼<choice>
    <reg>сидят</reg>
    <orig>сидятъ</orig>
</choice>
```

Fig. 2. An example of new and old spelling.

References

Andreev, A.V., Bukharkin, P.E., Matveev, E.M., Ponomareva, M.V.(2009) On development of new theoretical model for representation of literary history [O razrabotke novoi teoreticheskoi modeli representacii istorii literatury]. *Literaturnaia kul'tura Rossii XVIII veka* (Literary Culture of Russia of XVII century), 3, pp.303-310

Bonch-Osmolovskaya, A., Kolbasov, M., (2015) Tolstoy digital: Mining biographical data in literary heritage editions. *1st Conference on Biographical Data in a Digital World 2015*, *BD 2015;Amsterdam; Netherlands; 9 April 2015. CEUR-WS.org*, . pp. 48-52.

Burnard, L. (2014). *What is the Text Encoding Initiative?: How to add intelligent markup to digital resources*. OpenEdition Press.

Earhart, A. E. (2015). Traces of the Old, Uses of the New: The Emergence of Digital Literary Studies. *Editorial Theory and Literary Criticism*.

Lee, K. H., Slattery, O., Lu, R., Tang, X., & McCrary, V. (2002). The state of the art and practice in digital preservation. *Journal of Research-National Institute of Standards and Technology*, *107*(1), 93-106.

DCH-RP (2014), *Handbook: A Roadmap for Preservation of Digital Cultural Content*. Available at: dch.rp-eu (accessed 1 February 2015)

Mjør, K. J. (2009). The Online Library and the Classic Literary Canon in Post-Soviet Russia: Some Observations on "The Fundamental Electronic Library of Russian Literature and Folklore". *Digital Icons: Studies in Russian, Eurasian and Central European New Media*, *1*(2), 83-99.

Nerler P., (2014) CON AMORE Chapters from new book [CON AMORE Chapters from new book] *Sem' iskusstv* (Seven Arts), 4(51)

Opul'skaya L., (1997) The Academic Edition of Tolstoy's work., *Tolstoy Studies Journal,* 9., pp. 92-95

Orekxov B. V., Gallyamov A. A., Danilin S. Iu.Principles and goals of digital edtition of folklore archive (2012) [Principy i celi e'lektronnogo izdaniya fol'klornogo arxiva Bashkirskogo gosudarstvennogo universiteta] *Informacionnye texnologii i pis'mennoe nasledie: materialy IV mezhdunar. nauchn. konf. (Petrozavodsk, 3—8 sentyabrya 2012 g.)* (Informationa Technologies and Written Heritage). — Petrozavodsk, Izhevsk, pp. 198—201

Skorinkin, D.A. (2016) The TEI Standard as a universal instrument of digital editing [Standart TEI kak universal'nyi instrument e'lektronnogo predstavlenia teksta.] *Vestnik MGU* forthcoming

Robinson, P. (2013). Towards a theory of digital editions. *Variants: The Journal of the European Society for Textual Scholarship*, *10*, 105-31.

Votinsev P., (2006) The use of TEI Framework for data exchange in search-engine "Manuscript" [Ispol'zovanie formata TEI dlia obmena dannymi s polnotekstovoi informacionno-poiskovoi sistemoi «Manuskript»] Available at http://manuscripts.ru/conf/report/VotintsevP.pdf (accessed 1 February 2015)

Web Sites

Anthology of the 18th century Russian literature. Available at http://antology-xviii.spb.ru/

Decameron Web. Available at http://www.brown.edu/Departments/Italian_Studies/dweb/index.php

Emily Dickinson Archive. Available at http://www.edickinson.org/

Fundamental Electronic Library of Russian literature and Folklore (FEB). Available at http://feb-web.ru/

Henrik Ibsen's Writings. Available at http://www.ibsen.uio.no/

Mandelstam world. Available at http://www.mandelstam-world.info/

Manuscript. Available at http://manuscripts.ru/.

TEI Guidelines. Available at http://www.tei-c.org/Guidelines/P5/.

Transcribe Bentham. Available at http://www.transcribe-bentham.da.ulcc.ac.uk/

The Archive of Bashkir Folklore. Available at http://lcph.bashedu.ru/index.php?go=editions.

The Poetess Archive. Available athttp://www.poetessarchive.org.

Victorian Women Writers' Project. . Available at Vincent van Gogh – The Letters. Available at http://vangoghletters.org

http://webapp1.dlib.indiana.edu/vwwp/welcome.do.

World of Dante. Available at http://www.worldofdante.org/

Wright American Fiction 1851-1875. Available at http://webapp1.dlib.indiana.edu/TEIgeneral/welcome.do?brand=wright.

Цифровое издание произведений Льва Толстого: вклад в развитие российских литературных стипендиальных программ

А.А. Бонч-Осмоловская

Национальный исследовательский университет
«Высшая школа экономики»
Россия, 101000, Москва, ул. Мясницкая, 20

В статье рассматривается практика создания цифровых изданий в рамках сохранения цифрового наследия. Дается краткий обзор уровня технического развития в России. Утверждается, что современный проект по созданию цифрового издания должен отвечать ряду требований, касающихся долговременного хранения, доступности данных и устойчивости форматов. В статье освещается новый проект о литературном наследии Льва Толстого (Tolstoy Digital). Проект, в основании которого лежит 90-томное полное собрание произведений Толстого, тем не менее, позиционируется не как простое цифровое воспроизведение печатных изданий, а как самостоятельный ресурс, открытый для получения данных из других ресурсов.

Ключевые слова: цифровое издание, Лев Толстой, TEI, инициатива по кодированию текстов, цифровое сохранение.

Научная специальность: 10.00.00 – филологические науки.