Digital Edition of Complete Tolstoy's Heritage: OCR Crowd Sourcing Initiative, Literary Scholarship and User Scenarios

Bonch-Osmolovskaya, Anastasia
abonch@gmail.com
DH CLOUD; TOLSTOY DIGITAL

Orekhov, Boris
nevmenandr@gmail.com
DH CLOUD; Institute of Russian Literature (Pushkin House), Russia

Tolstaya, Fekla
6975991@gmail.com
TOLSTOY DIGITAL

The publication of the 90-volume complete Tolstoy's edition took thirty years (1928-1958). Despite the great effort put into the collection, the print run was small, making it a bibliographic rarity today. The edition contains more than 7.8 mln words, or 44,350 pages of extraordinarily diverse text, and consists of three parts: Fiction and Essays, including previously unpublished versions and drafts (volumes 1-45), Tolstoy's Diaries (volumes 46-58), Tolstoy's Correspondence (59-90). In 2014 the Tolstoy Museum in Moscow and IT company, ABBYY, a leader in OCR technology launched a unique crowdsourcing project "All Tolstoy in one click". The 90-volume edition was digitized using ABBYY's OCR technology and then proofread by thousands of volunteers from forty-nine countries within two weeks. The xml files that emerged from the crowdsourcing project gave birth to the idea of developing a fully-fledged digital edition of Tolstoy's heritage.

The crucial conceptual decision we had to start was whether we stick with the 90-volume publication as the material source, and create the digital version of the book, or we create a new digital edition of Tolstoy's heritage (Bonch-Osmolovskaya Skorinkin et al 2019). In this respect, digital output can pursue three goals, each of which has a direct influence on the final product:

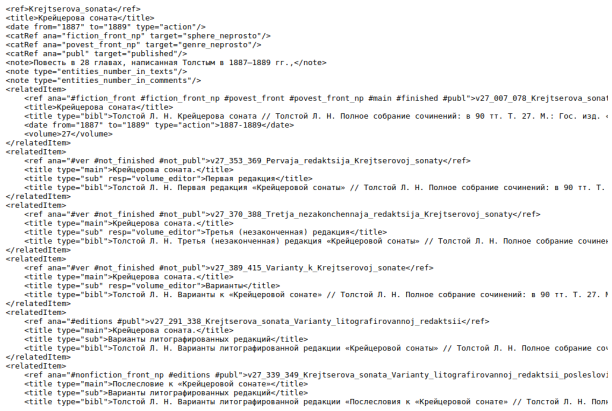
1. Preservation of Tolstoy's heritage, freed from the editorial construct of the 90-volume complete edition and open to further expansion by other sources.
2. Preservation of the enormous literary scholarship heritage, contained in the 90-volume edition in various critical apparatuses.
3. Accessibility of Tolstoy's heritage to the digital user, which ultimately means other scenarios of interaction with texts and their interpretations.

These same three goals were set by the creators of the "91st volume" application, which implements in electronic form one part of the complete works of Tolstoy, an index to the edition (Orekhov et al. 2018, Orekhov 2020).

The three goals are in fact in conflict, i.e. choosing only one of them would make it more difficult to fulfill the others. The first goal can be achieved, for example, by extracting only Tolstoy's text from the recognized OCR files. The second goal can best be achieved by creating a digital diplomatic edition (Pierazzo 2011). To achieve the third goal, one could, for example, skip the time-consuming phase of TEI preparation.

We have worked out a compromise that strikes a balance between the three goals and so far protects each of them through a series of specific conceptual choices.

1. We focus on Tolstoy's digital heritage with 90-volume edition being the primary, but not the only source. This means that we work with texts and not with volumes. We extract all of Tolstoy's text and provide each document with detailed metadata. We introduce the concept of a "family" of works, an abstract unit that groups together all variants of a text (Robinson 2013) and commentaries and other related works. The titles of all members of the family are converted into a machine-readable format, to between Tolstoy's original titles from those assigned that have been given by the editors of a volume. Each work in a family has a status tag that is used to create a hierarchy between the main work and the different types of its variants. Finally, families themselves can be linked if there is an important connection between them. The screenshot below shows an example.



```
<ref>Krejtserova_sonata</ref>
<title>Крейцерова соната</title>
<date from="1887" to="1889" type="action"/>
<catRef ana="fiction_front_np" target="sphere_neprosto"/>
<catRef ana="povest_front_np" target="genre_neprosto"/>
<catRef ana="publ" target="published"/>
<note>Повесть в 28 главах, написанная Толстым в 1887–1889 гг.,</note>
<note type="entities_number_in_texts"/>
<note type="entities_number_in_comments"/>
<relatedItem>
    <ref ana="#fiction_front #fiction_front_np #povest_front #povest_front_np #main #finished #publ">v27_007_078_Krejtserova_sonat
    <title>Крейцерова Соната</title>
    <title type="bibl">Толстой Л. Н. Крейцерова соната // Толстой Л. Н. Полное собрание сочинений: в 90 тт. Т. 27. М.: Гос. изд.
    <date from="1887" to="1889" type="action">1887-1889</date>
    <volume>27</volume>
</relatedItem>
<relatedItem>
    <ref ana="#ver #not_finished #not_publ">v27_353_369_Pervaja_redaktsija_Krejtserovoj_sonaty</ref>
    <title type="main">Крейцерова соната.</title>
    <title type="sub" resp="volume_editor">Первая редакция</title>
    <title type="bibl">Толстой Л. Н. Первая редакция «Крейцеровой сонаты» // Толстой Л. Н. Полное собрание сочинений: в 90 тт. Т.
</relatedItem>
<relatedItem>
    <ref ana="#ver #not_finished #not_publ">v27_370_388_Tretja_nezakonchennaja_redaktsija_Krejtserovoj_sonaty</ref>
    <title type="main">Крейцерова соната.</title>
    <title type="sub" resp="volume_editor">Третья (незаконченная) редакция</title>
    <title type="bibl">Толстой Л. Н. Третья (незаконченная) редакция «Крейцеровой сонаты» // Толстой Л. Н. Полное собрание сочине
</relatedItem>
<relatedItem>
    <ref ana="#ver #not_finished #not_publ">v27_389_415_Varianty_k_Krejtserovoj_sonate</ref>
    <title type="main">Крейцерова соната.</title>
    <title type="sub" resp="volume_editor">Варианты</title>
    <title type="bibl">Толстой Л. Н. Варианты к «Крейцеровой сонате» // Толстой Л. Н. Полное собрание сочинений: в 90 тт. Т. 27. М
</relatedItem>
<relatedItem>
    <ref ana="#editions #publ">v27_291_338_Krejtserova_sonata_Varianty_litografirovannoj_redaktsii</ref>
    <title type="main">Крейцерова соната.</title>
    <title type="sub">Варианты литографированных редакций</title>
    <title type="bibl">Толстой Л. Н. Варианты литографированной редакции «Крейцеровой сонаты» // Толстой Л. Н. Полное собрание со
</relatedItem>
<relatedItem>
    <ref ana="#nonfiction_front_np #editions #publ">v27_339_349_Krejtserova_sonata_Varianty_litografirovannoj_redaktsii_posleslov
    <title type="main">Послесловие к «Крейцеровой сонате»</title>
    <title type="sub">Варианты литографированных редакций</title>
    <title type="bibl">Толстой Л. Н. Варианты литографированной редакции «Послесловия к «Крейцеровой сонате» // Толстой Л. Н. Поли
```


2. We carefully include all critical apparatus, presented in the 90-volume edition. The printing practices of the mid-20th century involve certain reader scenarios that should be transformed into digital scenarios. For example, the index, which occupies a separate additional 91 volume, needs to be transformed into a database linked to a text span rather than the pages of a volume (Iglesia, Göbel 2015). The value of the index should not be underestimated: it cannot be replaced by a NER algorithm. The index reflects years of thorough editorial work aimed to clarify indirect mentions of people (Orekhov 2020). The 91 volume had not been proofread by volunteers and was, above all, full of OCR errors. So first we parsed and corrected the index, converted it to a database, assigned identifiers to each person entry and linked them to wikidata if possible. Secondly, we used SpaCy NER to extract all the persons mentioned in the documents and then automatically picked out the best candidates from the database. We ranked the candidates with similarity weights and then manually checked those whose weight was not 100%.
3. We have developed user scenarios to navigate the vast Tolstoy heritage. For example, we have simplified the genre sys-

tem to 8 basic categories out of 28 in the 90-volume edition, classified all persons from our index database with four parameters: Occupation, Lifetime (ancient time, Tolstoy's predecessor, Tolstoy's contemporary), Member of Tolstoy's family, Tolstoy's correspondent. We have developed a web interface with a full-search in Tolstoy works, databases, and a dictionary of rare words.

The essential idea that has guided our approaches and decisions is the idea of stages. The potential scope of the digital edition is so immense that we have been urged to plan our work so that each stage would lead to a finished product. The whole project consists of three stages: Digitalization, Creation of a digital scholarly edition, Creation an extended digital edition. The table below shows phases.

Table 1:

Phase	Stage name	Goal: Tolstoy's heritage	Goal: Literary scholarship	Goal: Digital scenarios
1	Digitalization of 90 volumes	- volumes to docs; - extract all metadata - basic tei markup,	- notes linking	
2	Digital scholarly edition	-complete tei mark-up -families of works	- index parsing - critical apparatus	- texts and person clasification - web-portal for search and navigation in texts and reference lists
We are here				
3	Digital edition of Tolstoy's heritage	Introduction of new documents from other sources: manuscripts from Tolstoy's museums and archives, audiofiles, images etc.	Extension of Tolstoy's bibliography from other sources	Development of new apps and websites on the basis of digitalized Tolstoy's heritage

All the tei documents are open access and can be found here https://github.com/tolstoydigital/TEI

Bibliography

Bonch-Osmolovskaya, Anastasia, et al . (2019): “ Tolstoy semanticized: Constructing a digital edition for knowledge discovery ”, in: *Journal of Web Semantics* 59: 100483.

Iglesia, Martin de la / **Göbel, Mathias** (2015): “ From entity description to semantic analysis: The case of Theodor Fontane's notebooks ”, in: *The Linked TEI: Text Encoding in the Web* 21.

Orekhov Boris / **Fischer, Frank** (2018): The 91st Volume-How the Digitised Index for the Collected Works of Leo Tolstoy Adds A New Angle for Research ”, in: *Digital Humanities 2018* . Puentes-Bridges. Book of Abstracts: 465-466.

Orekhov, Boris (2020): “ " Volume 91 " : An electronic index to the complete works of Leo Tolstoy ”, in: *Journal of Siberian Federal University. Humanities and social sciences.* 12: 2049-2055.

Pierazzo, Elena (2011): “ A rationale of digital documentary editions ”, in: *Literary and linguistic computing* 26.4: 463-477.

Robinson, Peter (2013): “ Towards a theory of digital editions ”, in: *The Journal of the European Society for Textual Scholarship* : 105-131.