Napis XXIV 2018
Doi: 10.18318/napis.2018.1.19


# Семантическое издание текстов Л. Н. Толстого: от текста к онтологии

Анастасия Бонч-Осмоловская, Матвей Колбасов, Борис Орехов, Ирина Павлова, Даниил Скоринкин
(Национальный исследовательский университет «Высшая школа экономики», Москва)

В этой статье мы расскажем о работе по созданию цифрового издания текстов Льва Толстого. Наша цель – появление многоуровневой разметки в большом и жанрово разнообразном собрании произведений русского писателя, но попутно мы хотели бы и внести свой вклад в стандартизацию электронного представления текстов и электронной текстологии в России. Мы считаем важным создавать образцы таких подходов для текстов не на английском языке, для которого уже есть довольно много авторитетных ориентиров. Что касается такого русскоязычного опыта, то на сегодняшний день он довольно беден и часто игнорирует стандарты цифровой публикации, такие, как Linked Open Data и Text Encoding Initiative. Мы надеемся, что наша работа позволит популяризировать эти стандарты, продемонстрировав их потенциал для выстраивания инфраструктуры сохранения и исследования текстов.

Наш основной источник – это 90-томное (т. н. «Юбилейное») *Полное собрание сочинений Льва Толстого*, выходившее в 1928-1958 годах. Это не просто печатное представление текста, за каждым томом стоит огромная текстологическая и исследовательская работа, которую нам также хотелось бы оцифровать и сохранить. Наша разметка должна сделать редакционные примечания более приспособленными для количественных исследований. Включенный в 90-томник критический аппарат, комментарии, биографические справки, указатели представляют собой фактически энциклопедическую базу знаний русской культуры с 1850-х до 1900-х годов. На сегодняшний день весьма небольшая часть этой базы представлена в открытых источниках вроде DBpedia [https://dbpedia.org].

Кроме того, большая часть нашей работы связана с подготовкой платформы по оцифровке и электронному представлению рукописей самого Толстого, а также других русских писателей. В этой статье мы дадим краткий обзор специального редактора для работы с такими источниками, системы «Текстограф».

Таким образом, наша деятельность на самом деле не сосредоточена только на фигуре самого Толстого, речь идет про создание определенной системы координат электронной публикации в России, а Толстой служит в какой-то мере удобным примером для конструирования издательских методов и практик. Он удобен в этом смысле по нескольким причинам. Во-первых, Толстой – известный (и даже более чем известный) писатель, а также основатель и лидер целого интеллектуального направления. Во-вторых, в его *Полном собрании сочинений* мы находим мало с чем сравнимое жанровое разнообразие от больших романов до дневниковых записей, включая сказки для детей, публицистические тексты, драматические произведения и другие с трудом поддающиеся категоризации примеры. В-третьих, Толстой обладал широкими социальными связями, и в поле его зрения попадали совершенно разные персоналии, так или иначе упомянутые в 90-томнике. В-четвертых, его тексты относятся к свободно распространяемым, не ограничены авторским правом и доступны в хорошем качестве в электронной форме. Эти обстоятельства делают подготовку научного цифрового издания сложной задачей и требуют применения широкого спектра новых редакционных практик и цифровых инструментов.

Нашей первой задачей была оцифровка богатого редакционно-критического аппарата печатного издания. Как мы ожидаем, семантическая разметка этих сведений превратит коллекцию цифровых текстов в базу данных с обширными поисковыми возможностями. Вторая задача – оцифровка именного указателя к 90-томнику, выделенного в специальную книгу (т. н. «91-й том») и представление его в виде отдельного сетевого ресурса. Указатель содержит более 16 000 имен и названий. Третья задача – выйти за рамки цифрового кодирования текста и предусмотреть полезные для исследователя литературы и историка культуры цифровые инструменты и способы подхода к дигитализованному материалу. Для решения этой задачи мы создаем специальный слой семантической разметки имен, речи персонажей и т. д.

Издание полного собрания сочинений Льва Толстого заняло 30 лет – с 1928 по 1958 год, а дополнительный том с указателями к собранию вышел в 1964 году. Несмотря на поистине титанические усилия, вложенные в это издание, тираж его был небольшим, и к настоящему времени оно уже стало библиографической редкостью. Собрание делится на три части. Тома с 1-го по 45-й содержат художественные произведения и публицистику (сюда же включены неопубликованные варианты и черновые редакции); в томах с 46-го по 58-й находятся дневники Толстого; тома с 59-го по 89-й – это письма Толстого. 90-й том – дополнительный, в него включены

произведения, которые не успели опубликовать в соответствующем томе по мере их выхода в печать.

В целом наследие Толстого содержит более 14,5 миллионов слов на 46 820 страницах. Для сравнения, тексты Ф.М. Достоевского, включенные в Национальный корпус русского языка, насчитывают всего 2 млн. слов.

Легенда гласит, что начало 90-томному собранию положил В.И. Ленин на заре 1920-х годов. Глава молодого советского государства отдал приказ опубликовать все, что было написано Толстым, чтобы человек новой формации получил доступ к наследию великого писателя. После подготовки и издания первых томов работа остановилась, так как некоторые толстовские тексты не могли быть пропущены советской цензурой. Это была своего рода патовая ситуация, поскольку, согласно воле Ленина, издание не могло быть остановлено, но и не могло продолжиться. В результате редакторские установки издания стали меняться: критический аппарат был сокращен, из издания были исключены письма корреспондентов Толстого, которые первоначально планировалось включить в многотомник. В таком измененном виде издание все же было осуществлено, хотя, как уже упоминалось, было малодоступным.

Ситуация изменилась в 2014 году, когда уникальный краудсорсинговый проект «Весь Толстой в один клик», инициированный Государственным музеем Л.Н. Толстого и фирмой ABBYY, привлек множество волонтеров из 49 стран, которые втечение двух недель исправили ошибки автоматического распознавания 90-томника, что в итоге позволило представить его в машиночитаемой форме. Сейчас тексты можно скачать с сайта tolstoy.ru в одном из форматов, используемых в программах для чтения с экрана. Одновременно с этим была подготовлена и xhtml-версия этих текстов. На этой xhtml-версии мы и выстраиваем наше семантическое издание текстов Л.Н. Толстого. Размеченные тексты размещены в открытом доступе по адресу [https://github.com/tolstoydigital/TEI].

Первое стратегическое решение, которое мы должны были принять в ходе работы над нашим изданием, – это фокусировка на произведениях Толстого, а не на их представлении в 90-томном издании. Это позволило нам отделить собственно наследие писателя от идеологически нагруженных предисловий советских редакторов. Другие разновидности критического аппарата (редакционные пометы, комментарии, послесловия) были проанализированы и разделены на две группы: то, что можно представить в виде текстовой разметки, и то, что нужно представить в виде метаданных к произведению.

Оригинальное издание складывается из томов как основной структурной единицы. Мы сосредоточились на отдельных произведениях, что заставило нас сегментировать печатное издание на самостоятельные тексты (романы, рассказы, сказки, статьи, письма и т. д.).

Поскольку 90-томник состоит из трех больших классов текстов (художественные и нехудожественные произведения, дневники и письма), каждый из них требовал особенного подхода при создании соответствующего цифрового объекта.

В процессе работы мы разделили все тома, содержащие произведения Толстого, но не стали проводить такого же членения для томов, содержащих дневники. Причина этого в том, что критический аппарат жестко привязан к тому набору дневниковых записей, которые содержатся в данном томе. Однако внутри тома дневники сгруппированы по тетрадям, отражающим связь с физическим источником этих текстов, и эта группировка также заложена в нашей разметке.

Первый уровень разметки находится в области метаданных. Мы полуавтоматически создали базу данных 90-томника, в которой каждый текст связан с набором атрибутов. В первую очередь, тексты были классифицированы на три основных типа: письма, дневники и произведения. Тексты произведений затем подверглись более дробному делению на художественные, нехудожественные (публицистические), религиозные и детскую литературу. На этом уровне мы также приписали каждому тексту уникальные названия (такие, которые даны им в 90-томнике). Каждое письмо получило искусственный «псевдоним», сложившийся из имени адресата и даты (как они представлены в 90-томнике). Дневники именованы по датам. Художественные и публицистические тексты Толстого также были классифицированы по жанрам. В этой классификации мы опирались на редакторское описание в печатном издании. Но для удобства пришлось пойти на некоторые упрощения: исходная система жанров включала большой массив типов от романов, рассказов, пьес до сказок, школьных учебников, молитв, текстов публичных выступлений, завещания и перевода Евангелия. Некоторые из этих типов имели всего одно представление в корпусе.

Публицистические произведения были классифицированы по тематике: «искусство», «философия и религия», «образование», «медицина», «политика и общество». В отличие от жанрового деления, публицистика могла иметь множественную тематическую привязку (например, «политика» и «религия» для одного и того же текста). Эта классификация позволяет оценить эволюцию Толстого как публициста. Журналистская активность Толстого впервые проявляет себя в 1860-х, когда писатель поселяется в Ясной Поляне и начинает свою образовательную деятельность. Он открывает свою собственную школу для крестьянских детей и пишет о своем опыте преподавания. Главная тема этого времени – образование. В 1880-х годах, сразу после пика «морального кризиса» Толстого, писатель становится влиятельной общественной фигурой. Его главные нехудожественные темы – это политика и общество. Наконец, в последние годы жизни Толстой все больше времени и внимания посвящает теме религии и собственному духовному опыту. После побега из Ясной Поляны Толстой напишет около 30 фрагментов, которые целиком посвящены именно религии.

Для всякого критического издания чрезвычайно важно сохранить информацию о статусе текста. В нашем используется два бинарных атрибута: «закончен» и «опубликован». Любой незаконченный фрагмент, набросок или черновик будет помечен как незаконченный и неопубликованный. Завершенная версия любого текста, которая прошла процесс публикации до того, как началась работа над 90-томником, помечается как законченная и опубликованная. Но кроме этих двух типов возможны промежуточные случаи. Разные варианты одного текста помечаются в специальном атрибуте метаданных.

Некоторым текстам заголовки даны издателями собрания сочинений. Эти заголовки обычно содержат либо первую строку текста, либо краткое его описание. Так как работа над изданием шла несколько десятков лет, сохранить единообразие заголовков было невозможно. Для таких отрывков и набросков мы предложили собственную структуру названия. Это позволило нам группировать тексты при автоматическом обращении к электронной коллекции.

Сокращенный вариант TEI-заголовка документа мы приводим ниже:

```
<teiHeader>
  <fileDesc>
...
   <sourceDesc>
    <biblStruct>
     <analytic>
      <author>
      Толстой Л.Н.
      </author>
      <title level=»a»>
      «Варианты к статье «О Шекспире и о драме»»
      </title>
      <title type=»normalized»>
         <metatitle>
         ”О Шекспире и о драме»
         </metatitle>
         <version=”variant”>
         Варианты
         </version>
         <version_number>undefined </version_number>
      </title>
     </analytic>
     <monogr>
```

```
      <title level=»m»>
      Полное собрание сочинений. Серия первая «Произведения». Том 35
      </title>
     </monogr>
    </biblStruct>
   </sourceDesc>
  </fileDesc>
...
  <profileDesc>
   <creation>
    <date notAfter=»1904» notBefore=»1903»>
    1903-1904
    </date>
...
   </creation>
   <textClass>
    <catRef target=»#публицистика» type=»sphere» />
    <catRef target=»#публицистика.статья» type=»type» />
    <catRef target=»#искусство, философия и религия» type=»domain»/>
   </textClass>
   <preparedness>
   not finished
   </preparedness>
   <isEdited>
   not edited
   </isEdited>
  </profileDesc>
...
</teiHeader>
```

Итак, принципиальной целью семантического издания текстов Толстого было извлечение релевантных метаданных из критического аппарата и приведение их к формату укороченного списка.

Особого подхода потребовала разметка писем и дневников.

Во-первых, их можно было размечать в большой мере автоматически. Собрание сочинений Толстого содержит 31 том писем, датированных с 1841 по 1910 года. Все письма представлены в хронологическом порядке. В собрании 7 томов (с 83-го по 89-й) писем, адресованных только одному человеку: либо Владимиру Черткову, близкому другу Толстого, либо жене писателя Софье Андреевне Толстой.

После разделения всех томов на отдельные документы получилось 10 529 писем. Каждому письму сопоставлена информация об адресате, дате и месте написания, а также сведения о статусе письма (было ли оно отправлено, является ли оно черновиком). Кроме того, письмо сопровождается детальным комментарием, содержащим биографическую информацию об адресате. Таким образом, наши действия были предопределены, с одной стороны, структурой текста в печатном издании (точнее говоря, распределением информации между текстом письма, сносками и ссылками, а также комментарием, следующим за текстом письма), с другой стороны, доступными в TEI-схеме элементами для кодирования. Сведения о письме были извлечены автоматически и помещены нами в заголовок документа. Туда же была помещена информация о публикации (вроде «публикуется впервые»). Всего заголовок содержит около 10 атрибутов, описывающих письмо, значения которых были автоматически извлечены из следующего за текстом комментария. Даты были унифицированы в соответствии со стандартами TEI. Отдельную проблему составила нормализация имен адресатов. Всего в издании размещены письма к 2800 уникальным адресатам.

Конвертирование данных о дневниковых записях в TEI-представление связано частично с теми же трудностями, что возникают и при работе с письмами: разнообразие форматов записи дат. Дело усложняется тем, что многие записи имеют не только авторскую, но и редакторскую датировку. Это обстоятельство требует особенного внимания и добавления специального атрибута «resp» в теге <date>, который указывает на источник сведений о датировке.

Оформленные в TEI-представлении тексты уже сейчас доступны для скачивания и автоматической обработки. Но в будущем планируется создать на их основе специализированный сайт с гибкими возможностями поиска. Механизмы для полнотекстового поиска хорошо адаптированы для русского языка. Однако и тут есть нюанс. Русский язык обладает богатой словоизменительной морфологией, и чтобы пользователь смог найти нужное слово, его обычно приводят к словарной форме с помощью специальных программ – леммеров. Однако эти программы умеют работать только с текстами, представленными в новой орфографии, утвержденной для русской графики с 1918 года. Часть текстов Л.Н. Толстого издана в 90-томнике в оригинальном виде в старой орфографии (прежде всего, это касается черновиков, редакций и вариантов). В TEI есть пара специализированных тегов, с помощью которых мы можем отразить оба написания: <orig> и <reg>. Первый заключает в себя оригинальное написание, а второй – отредактированное (в нашем случае – конвертированное в новую орфографию). Специально для того, чтобы произвести эту операцию автоматически, была разработана система, переводящая тексты из старой орфографии в новую (она представлена в открытом доступе: https://github.com/shelari/prereform_to_contemporary). В эти теги были заключены те фрагменты

текста, которые изданы в старой орфографии. Для поисковой индексации будут использованы варианты, заключенные в тег <reg>.

Отдельный важный источник информации представляет собой 91-й том, объединяющий в себе указатели к каждому конкретному тому собрания. Самая ценная его часть – указатель имен собственных, содержащий более 16 000 имен. Этот указатель был конвертирован в формат базы данных, на основе которой построено и доступно для работы веб-приложение: [http://index.tolstoy.ru/]. В составе его функциональности – обратный индекс, позволяющий устанавливать факты совместной встречаемости имен на страницах издания. Этот факт может стать материалом для популярного сейчас в гуманитарных науках сетевого анализа.

Обычно этот метод ассоциируется с социальными сетями – известными сервисами для организации социальных взаимоотношений. Но сам метод шире и сложнее, с его помощью анализируются и посторонние по отношению к социальной жизни предметы (например, лингвистические или литературоведческие данные).

Суть метода в том, что он представляет в виде модели некоторые сущности (в данном случае, персоналии из указателя) и связи между ними. Сущность предстаёт в модели узлом графа, а связь – его ребром. Эта модель весьма перспективна для вычисления наиболее значимых узлов сети (например, с применением алгоритмов расчёта центральности) и распределения узлов по разным «подсетям», кластерам внутри большого графа.

*91-й том* развёртывает перед нами социальную сеть Л.Н. Толстого, при этом фиксируются связи не только Толстого с другими людьми (это можно видеть, например, по переписке), но и связи людей между собой. В основе этой сети лежит тот же указатель: факт совместной встречаемости каких-то имён на одной странице в 90-томнике становится основанием для прочерчивания ребра графа, то есть реализуется как связь внутри сети. Скажем, индийская *Бхагават-гита* встречается на страницах *Полного собрания сочинений* 5 раз, при этом на тех же страницах в общей сложности упоминается ещё 43 имени, и имена эти, разумеется, не случайные, они все так или иначе относятся к одной теме и образуют, например, «индийский кластер»: *Гитопадеша*, *Дхаммапада*, *Вамана Пурана*, *Рамакришна Шри Парамагамза*. Но Толстого эти тексты интересуют в том числе и как носители философского знания, поэтому связаны с ними оказываются и сугубо философские персоналии: *Вестник теософии*, Ксенофонт, Монтень, Монтескье, Паскаль, Сковорода, Сократ.

Эти сети дают большие возможности для охвата научными методами всего многообразия интересов и идей Толстого. Оценить размах можно на странице, изображающей всю сеть имён собственных. На ней представлена панорамная картина, которая даёт представление только об общих тенденциях и самых больших тематических кластерах. Но для каждого отдельного имени есть и свой маленький граф, показывающий наиболее значимые связи имён между собой.

Указатель, транслированный в формат открытых связанных данных (Linked Open Data), наглядно представляет интересы писателя и его интеллектуальную жизнь.

Важная параллельная работа ведется с цифровым представлением рукописей Толстого. Для технического обеспечения этой работы нами разработана специальная онлайн-система «Текстограф» (функциональность доступна по паролю на сайте textograf.ru). В рамках этой системы возможна загрузка отсканированного изображения рукописи, набор ее транскрипции, разметка содержащихся в рукописи правок, установление соответствий между частями транскрипции и местом их расположения на листе (mapping), а также более сложные операции, связанные с традиционной текстологической работой, подразумевающей установление аутентичных редакций произведений. Выгрузка размеченных данных производится в формате TEI.

Кроме работы с каждым отдельным листом рукописи, которая так или иначе выполняется при оцифровке рукописи любого автора, в «Текстографе» есть и специализированная функциональность, учитывающая интересы текстологов, работающих с авторами больших нарративных произведений, в том числе, Толстого. Речь идет о работе по формированию редакций, складывающихся из нескольких вариантов отдельных сцен. Для этого в «Текстографе» есть специальный интерфейс, напоминающий «монтажный стол» в кинематографической студии. Кажется, что такая функциональность (как и специализированная «карта источников», граф отдельных документов, содержащих планы, наброски и законченные варианты сцен) уникальна среди текстологических программ (Transkribus, Smart-GS и т. д.) и нигде, кроме «Текстографа», не представлена.

В будущем эти глубоко размеченные рукописи будут интегрированы в семантическое издание.

Так мы двигались в направлении создания большой онтологии текстов Толстого, в которой своей разметкой обладает каждый отдельный документ, а с текстом документа связаны элементы указателя. Вся эта работа должна стать основой для будущего цифрового исследования текстов русского писателя современными методами digital humanities, а сама текстовая разметка будет расширяться и в будущем покроет и отдельные понятия, свойства и отношения, упомянутые в романах и эго-документах Льва Толстого.

**Ключевые слова**: цифровое издание, TEI, семантическая разметка, Толстой

Anastazja Boncz-Osmołowska, Matwiej Kołbasow, Boris Oriechow, Irina Pawłowa, Daniił Skorinkin

## Semantyczna edycja tekstów Lwa Tołstoja: od tekstu do ontologii

W artykule przedstawiono rezultaty projektu cyfrowej edycji wydania *Dzieł zebranych* Lwa Tołstoja, opracowanego jako materiał w otwartym dostępie. Naszym głównym źródłem jest dziewięćdziesięciotomowe krytyczne wydanie tekstów Lwa Tołstoja. Praca ukazuje tworzenie struktury metadanych do tekstów, podzielonych na trzy grupy: teksty literackie, dzienniki i listy. Znaczniki pochodzą z aparatu krytycznego wydania, zaś samo oznaczanie pozwala na stworzenie obrazu ewolucji Tołstoja jako pisarza. Indeks do wydania krytycznego również stanowi ważne źródło danych, które zostało w ramach projektu cyfrowego zdigitalizowane i opracowane w postaci wyspecjalizowanej usługi sieciowej. Dane te pozwalają w szczególności na stworzenie sieci odniesień między istotnymi dla Tołstoja osobami i tekstami, a sam układ odniesień może stać się podstawą do analizy sieciowej, metodologii popularnej we współczesnych naukach humanistycznych. W końcowej części artykułu omawia się „Tekstograf" – platformę techniczną służącą do ucyfrowienia rękopisów, która powstała w celu zdigitalizowania rękopisów Tołstoja, ale może zostać wykorzystana również do prac nad manuskryptami innych pisarzy.

**Słowa klucze:** edycja cyfrowa, TEI, oznaczanie semantyczne, Lew Tołstoj

Anastazja Boncz-Osmołowska, Matwiej Kołbasow, Boris Oriechow, Irina Pawłowa, Daniił Skorinkin

## Semantic editing of the texts by Leo Tolstoy: from text to ontology

The article presents the results of a digital editing project devoted to Leo Tolstoy's *The Collected Works*, developed as an open access material. Our main source is the ninety-volume critical edition of Leo Tolstoy's texts. The article portrays the developing of a metadata structure for the texts, which were divided into three categories: literary texts, diaries, and letters. The markers derive from the mechanism of the critical edition, and the mark-up itself allows for a creation of an image of Tolstoy's evolution as a writer. The critical edition's index also constitutes an important source of data, which was digitised as part of the project and developed as a specialised web service. This data supports, in particular, the construction of a network of references between the people and texts essential to Tolstoy. Further-

more, the layout of references in itself might become the basis of social network analysis, a methodological technique, popular in contemporary humanities. In the final part of the article, the 'Textograph' is discussed – a technical platform serving to render manuscripts digital, which was developed in order to digitise Tolstoy's manuscripts, but can also be used to work on manuscripts of other writers.

**Keywords:** digital editing, TEI, semantic mark-up, Leo Tolstoy