# ЧТО ТАКОЕ СЕМАНТИЧЕСКОЕ ИЗДАНИЕ И ПОЧЕМУ В БУДУЩЕМ ВСЕ ИЗДАНИЯ СТАНУТ СЕМАНТИЧЕСКИМИ?

MIKHAIL GRONAS, BORIS OREKHOV (Hanover / Moscow)

Seid umschlungen, Millionen! J. C. F. von Schiller, An die Freude

Наша наука чаще всего занимается прошлым, настоящим реже, а будущим почти никогда.

А Александру Константиновичу Жолковскому, как знают его коллеги, ученики и читатели, присуща интеллектуальная чуткость и неувядающий юношеский интерес к самоновейшему в культуре, к тому, что происходит прямо сейчас или вот-вот произойдет. Поэтому мы решили посвятить ему статью-предсказание о будущем филологии и литературоведения. Статья состоит из двух частей: в первой содержится наше пророчество, которое, в соответствие с законами этого жанра, темно и загадочно. А во второй мы попытаемся его истолковать.

I

В будущем все научные издания станут семантическими. А потом они сольются в одно — единственное и единое — семантическое

*издание* (или *семантическую сеть*) всего культурного наследия человечества.

#### П

## От цифровых текстов к семантическим

Очевидно, что формы существования текстов зависят от принятых в обществе технологий хранения и воспроизводства информации. В бесписьменных культурах главный носитель информации — индивидуальная и коллективная память, соответственно и тексты, которые эти культуры стремятся сохранить, существуют как объекты, приспособленные к мнемоническому выживанию. Формальная организация фольклорных текстов, например описанные Лордом и Пэрри формулы эпоса, или пропповский набор функций в волшебной сказке — это элементы мнемонической технологии, которые соотносятся с памятью как паровой двигатель с паром.

На смену памяти пришли письмо и печать — технологии, которые привели к появлению рукописи и книги. Изобретение и распространение компьютера вызвало к жизни цифровой текст. Когда компьютеры объединились во всемирную сеть и научились обмениваться документами при помощи гиперссылок, основной формой существования текста стал гипертекст. А уже в конце 90-х изобретатель протокола обмена гипертекстами (и, стало быть, Интернета) сэр Тим Бернерс-Ли заговорил о том, что придуманная им Всемирная Паутина устарела, а на смену ей придет новая семантическая (semantic web) $^1$ . По мысли Бернерса-Ли сегодняшняя сеть, то есть сеть гипертекстов — «человеческая, слишком человеческая»; она сплетена из документов, созданных людьми и для людей. Машина (например, поисковая машина), конечно, очень хорошо ищет нужные нам слова, но больше, в сущности, помочь ничем не может, потому что Интернет (как ни парадоксально это звучит) немашиночитаем. А грядущая Семантическая Сеть свяжет уже не поверхностные представления текстов, а стоящие за ними

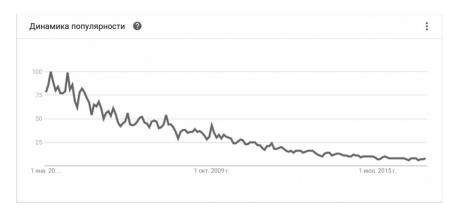
<sup>1</sup> Cm.: Tim Berners-Lee, "Semantic Web Road map" [October 14, 1998], URL: https://www.w3.org/DesignIssues/Semantic.html (accessed April 14, 2017); Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", *Scientific American* 284.5 (2001): 34–43. URL: https://www.scientificamerican.com/article/the-semantic-web/ (accessed April 14, 2017).

смыслы и данные, и откроет их для автоматического анализа. От поиска слов мы перейдем к пониманию причин, следствий, связей и отношений между вещами, явлениями, процессами, потоками данных. Семантическая сеть станет новой формой существования текста, в которой будут наконец реализованы возможности новых носителей информации, компьютера и компьютерных сетей. Для воплощения семантической утопии нужно было всего лишь научиться представлять документы в машиночитаемом виде, т.е. в виде упорядоченных и универсально распознаваемых структур данных. Бернерс-Ли и его соратники разработали довольно простые инструменты, при помощи которых обычные документы можно превратить в семантические. Каждой сущности (ресурсу) приписан универсальный (т.е. неизменный, универсально распознаваемый) идентификатор (URI). Области знаний и данных описываются в виде иерархически организованных онтологий. Отношения и свойства сущностей (ресурсов) задаются набором простых утверждений (субъект/предикат/объект), Resource Description Framework. Если добавляемая в Интернет информация будет организована подобным образом, она, наконец, оживет, данные заговорят друг с другом, цифровые тексты станут семантическими, обещал Бернерс-Ли в 1998 году.

Прошло почти 20 лет, а семантическая утопия сэра Тима так и не наступила<sup>2</sup>. Сам термин semantic web воспринимается теперь как чересчур амбициозный, и, как видно из графика популярности соответствующего поискового запроса в Google Trends, постепенно выходит из моды (см. рис. 1).

Главная причина неудачи проекта состоит в известном парадоксе tragedy of commons: для преобразования обычных документов в семантические требуется затратить время и усилия; а обещанное общее благо («семантическая утопия») — недостаточная мотивация для большинства авторов и публикаторов, которых вполне устраивает досемантический Интернет.

<sup>2</sup> Тем не менее, об идее семантического Интернета еще помнят: Abraham Bernstein, James Hendler, Natalya Noy, "A New Look at the Semantic Web", Communications of the ACM 59.9 (2016): 35-37, DOI: 10.1145/2890489 (accessed April 14, 2017). Бернерс-Ли в последние годы предпочитает говорить о связанных данных (Linked Data).



**Рис. 1** Динамика популярности понятия «semantic web» в поисковых запросах Google, https://goo.gl/O2Tzs9

При этом очень много отдельных элементов семантической сети реализованы<sup>3</sup>, и мы пользуемся ими ежедневно, когда обращаемся к упорядоченным и связанным структурам данных — например, сравнивая цены в интернет-магазинах, покупая авиабилеты, или когда одним кликом отправляем информацию о событии (приглашение пообедать, поход к доктору) из электронного письма в календарь. Иными словами, семантическая сеть прижилась в тех случаях, когда такая организация информации выгодна и полезна ее владельцу или публикатору.

Наш тезис состоит в том, что культурное наследие человечества, в частности литературные тексты, также относится к зоне потенциального применения семантической сети. Прежде всего потому, что речь идет о текстах, которым наша культура приписывает особую ценность; стало быть ценными (эстетически, исторически, и т.д.) для нас являются и стоящие за этими текстами смыслы и знания, и мы готовы ради них потрудиться, т.е. потратить усилия на упорядочивание данных. Филологи и историки культуры (так же как продавцы авиабилетов) заинтересованы в превращении вверенной им области знания в семантическую сеть.

<sup>3</sup> Fabio Ciotti, Francesca Tomasi, "Formal Ontologies, Linked Data, and TEI Semantics", *Journal of the Text Encoding Initiative* 9 (2016), URL: http://jtei.revues.org/1480 (accessed April 14, 2017).

Мало того, мы (наша цивилизация и наша профессия) этим уже давно занимаемся. Комментарии, индексы, выявление цитат, параллельных мест, установление авторства и датировок, функций у В.Я. Проппа или инвариантов у А.К. Жолковского и Ю.К. Щеглова — все эти древние и новые филологические практики почти ничем не отличаются от составления семантической сети. Недаром для обозначения совокупности упорядоченных сведений об издании мы пользуемся термином критический аппарат. Аппарат синоним машины; от него один шаг до машиночитаемости и семантических изданий.

### Что такое семантическое издание?

Семантическое издание — это представление текста в виде связанных данных. Для этого текст должен быть размечен, т.е. его элементам приписана информация об их значении в виде организованных меток в машиночитаемом формате. Притом формат должен быть универсальным: т.е. для разметки одних и тех же сущностей в разных текстах должны использоваться одни и те же метки. Универсальная разметка связывает данные разных текстов в предсказанную Бернерсом-Ли семантическую сеть; в этом смысле никакое семантическое издание не существует само по себе — no semantic edition is an island. Поэтому в пределе предметом семантического издания является не отдельный текст или собрание текстов, а вся совокупность написанного (и изобретенного, изображенного, сочиненного, сыгранного и т.д.) человечеством.

Что значит представить текст в виде данных? Данные — это набор отношений. Данные выражают прямую, эксплицированную и поэтому понятную компьютеру взаимосвязь сущностей. Скажем, если мы целый день мерим температуру воздуха, у нас могут получится вот такие данные: 8:00 — +18°C; 9:00 — +19°C; 10:00 — +20°С ... 16:00 — +26°С; 17:00 — +26°С. Здесь время увязано с уровнем температуры, связь эта обозначена графически, с помощью тире. Символы обозначающие участников отношения — время и температуру — единообразны и отграничены друг от друга: между часами и минутами стоит двоеточие; перед значением температуры стоит плюс (дело было летом), а после — кружок,

означающий «градус». Каждое наблюдение отделено от остальных с помощью точки с запятой, а час и градусы разведены тире. Эту последовательность легко представить в виде таблицы:

время	8:00	9:00	10:00	16:00	17:00
температура	+18°	+19°	+20°	+26°	+26°

Легко и написать программу, которая прочтет эту последовательность символов, разложит наши сущности по ячейкам памяти и проведет их количественный анализ (например, построит регрессионную модель и предскажет изменение температуры в будущем). Текст, даже самый прозрачный и очевидный, такого комфорта не предоставляет.

Вместо списка или таблицы мы могли бы написать:

Проснулся в 8. Вставать не хочется. Температура +18. Провалялся час в постели. За это время температура выросла на один градус. Вставать все еще не хочется. Еще час провалялся! И снова на градус! etc...

Получилась изящная дневниковая виньетка. Но компьютеру она не по плечу: сущности не помечены и не отграничены друг от друга, установить связи между ними невозможно.

Помочь компьютеру можно при помощи *разметки*. Она могла бы выглядеть так:

Проснулся в 8. [время, 8:00] Вставать не хочется. Температура +18. [температура, +18°C] Провалялся час в постели. [время, 9:00] За это время температура выросла на один градус. [температура, +19°C] Вставать все еще не хочется. Еще час провалялся! [время, 10:00] И снова на градус! [температура, +20°C]

#### O PA3METKE

Разметка — главный инструмент семантического издания. Разметкой называют систему нотации, которая позволяет компьютеру отличать одни сущности от других. Именно она и делает текст машиночитаемым. Сущности, о которых идет речь, бесконечно разнообразны — это все возможные виды данных (знаний, смыслов и т.д.), содержащиеся в тексте в любом виде (эксплицитно или имплицитно) и на любом уровне. Человек без труда понимает их смысл и отличает их друг от друга, машине же для этого требуется разметка. Зато с уже размеченными сущностями машина может делать то, что не по силам уже человеку: находить их мгновенно в миллионах текстов, обнаруживать связи и зависимости с другими данными, описывать статистические закономерности и исторические тенденции.

Простейший и самый распространенный пример разметки структурная разметка документа. Человек без труда различает заголовок, основной текст, примечание и заметку на полях. А для компьютера всё это просто последовательности символов. Разметка (в данном случае теги языка HTML) подсказывает, что вот эта часть текста обладает особым качеством — является заглавием (тег <h1>), а эта — абзацем (тег ). Браузер интерпретирует эти теги и превращает их в формальные элементы веб-страницы.

В основе современных лингвистических корпусов лежит лингвистическая разметка: сложные запросы, которые мы можем задавать к поисковым механизмам этих корпусов, также в конечном счете могут быть успешны только потому что в тех коллекциях, по которым мы ищем, нужные нам сущности были заранее выделены, то есть размечены. Так, базовым свойством современных корпусов является морфологическая разметка, то есть приписанные каждому слову определенные свойственные данной форме грамматические характеристики: часть речи, словоизменительный класс, место в словоизменительной парадигме. Всё это дает возможность «понимать», ОТР перед ним существительное, прилагательное, глагол или что-то еще. Такая разметка делается, в основном, автоматически, что позволяет составлять корпуса из мощных текстовых коллекций, достигающих десятков миллиардов словоупотреблений.

Разумеется, мы хотим объяснить компьютеру не только грамматику, но и смыслы — все, что понимает и может увидеть в тексте человек. Для этого нужна семантическая разметка — система нотации, которая выделяет в тексте сущности, отграниченные друг

от друга не формально, а содержательно. Сюда входит разметка лексико-семантических групп, например глаголы движения, названия животных, растений, термины родства, еда и напитки, предметные онтологии, социальные отношения и т.д.

Чтобы передать компьютеру не только смыслы, но и знания, заключенные в текстах, понадобится энциклопедическая разметка. Она указывает на типы данных традиционные выделяемые в критических аппаратах и указателях: именованные сущности, персоналии (авторы, адресаты, персонажи, исторические лица и т.д.), названия изданий, произведений; географическая информация (топонимы) и хронология (даты, упоминаемые в текстах, и датировки произведений).

Наконец, *историко-культурная* разметка выделяет в тексте более сложные сущности, например, литературные жанры, аллюзии на исторические события, следы влияния идей, школ или авторов, цитаты и т.д. Сюда же входят такие подвиды разметок, как, например, текстологическая (происхождение и описание источника текста, отношения между версиями текста), риторическая разметка (риторические фигуры), стиховедческая разметка, и разметка параллельных текстов (сведения об отношениях переводных эквивалентов).

Из этого списка ясно, что огромная часть того, чем занимаются лингвисты, филологи и литературоведы, может быть превращено в разметку. Список разметок принципиально открыт и потенциально бесконечен, поскольку бесконечно количество сущностей, которые могут когда-либо кого-нибудь заинтересовать. Если вас интересуют инфинитивы в русской поэзии, анаграммы у Мандельшатама, прически героев Толстого, скрытые цитаты из Монтескье у Джойса, институты дуэли или бала в русской классике, палиндромы в древнекитайских хрониках, карлики у Вячеслава Иванова, колики у Всеволода Иванова, что угодно, где угодно, когда угодно, на каком угодно языке, под любым углом зрения, в контексте любой методологии — вам ничто не мешает добавить соответствующую разметку к семантическому изданию интересующего вас текста или группы текстов. Нужно только позаботиться о том, чтобы эта разметка стала универсальной, т.е. соотнести ее с общим реестром всех существующих разметок. Это последнее действие гарантирует связанность данных, и через них связанность всех семантических изданий.

## Откуда возьмется разметка?

Во-первых, филологи будущего (или электронные издатели, грань между этими профессиями будет зыбкой) будут сами добавлять разметку в семантические издания. Результаты филологических и литературоведческих исследований часто можно представить в виде разметки. Любой добросовестный исследователь при публикации неизбежно уделяет место так называемой «метаразметке», информации, относящейся ко всему тексту сразу: когда, где, кем. Электронные издатели уже и сейчас довольно часто размечают структуру публикуемых текстов и текстологические элементы.

Во-вторых, поскольку количество возможных разметок ничем не ограничено, тексты могут размечаться и сторонними участниками. Уже сейчас одной из тенденций развития digital humanities становится привлечение дополнительных ресурсов через «crowdsourcing», в основе которого лежит уверенность в том, что можно достичь впечатляющих результатов, если за решение некоторой задачи возьмется достаточное число неравнодушных энтузиастов, не обязательно принадлежащих высококвалифицированному сообществу<sup>4</sup>. Семантические издания будут включать инструменты для коллективной и распределенной разметки<sup>5</sup>.

В-третьих, в руках человека оказывается всё инструментов автоматической разметки. Уже довольно давно

<sup>4</sup> Вот несколько недавних примеров. Большую помощь в оцифровке 90-томного собрания сочинений Л.Н. Толстого оказали включившиеся в процесс «простые» люди, исправлявшие ошибки распознавания: http://tolstoy.ru/projects/tolstoyin-one-click/ (accessed April 14, 2017). Набором рукописных дневников для проекта «Прожито» (см. ниже) также занимаются люди, вовлеченные в эту деятельность благодаря социальным сетям, а не профессиональной занятости. Заинтересованные пользователи, просмотревшие большое число фотографий поверхности Марса, помогли обнаружить место падения советской автоматической станции «Марс-3» (https://geektimes.ru/post/175827/ [accessed April 14, 2017]). Немного более специальный пример (потому что требующий от участников особой подготовки): сайт «Ancient Lives» (https://www.ancientlives. org/ [accessed April 14, 2017]), привлекающий волонтеров для трансрибирования и электронного издания греческих папирусов.

<sup>5</sup> Такие программные инструменты есть уже сейчас, например, разработанная в MIT программа для коллективного редактирования и комментирования: http:// www.annotationstudio.org/ (accessed April 14, 2017).

лингвисты и инженеры работают с программами-лемматизаторами, присваивающими каждой словоформе в тексте информацию о ее нормализованном виде (лемме). Аналогичные инструменты стали появляться и для других уровней организации текста, в частности, парсеры определяющие морфологические и синтаксические свойства лексем и семантические парсеры. Методы машинного обучения и статистического анализа языка могут применяться для более тонких разметок, например для нахождения тем/топиков характеризующих разные фрагменты текста.

Наконец, в-четвертых, за свою двухтысячелетнюю историю филология накопила огромные массивы сведений, содержащихся в комментариях, списках параллельных мест, примечаниях, указателях к бумажным изданиям и в разного рода библиографических справочниках, и т.д. Все аппараты традиционных научных изданий и так называемая справочная литература уже являются упорядоченными данными. Их осталось только дооцифровать и связать едиными стандартами разметки<sup>6</sup>. Семантические издания будут опираться на инструменты автоматического перевода данных из формата критических аппаратов прошлого в форматы современной стандартизированной разметки.

# Семантическая сеть культуры

Попробуем обобщить сказанное в виде схемы, передающей архитектуру (т.е. общее устройство) семантической сети культуры (см. рис. 2). Эта сеть состоит из связанных семантических изданий. Семантическими издания становятся благодаря разметке. Источники разметки: работа филологов издателей и филологического сообщества, сообщество читателей в целом (crowdsourced community), инструменты автоматической разметки, автоматическая обработка и интеграция уже существующих данных (критических аппаратов). Список разметок открыт (т.е. можно добавлять новые разметки), но унифицирован; существует единообразный стандарт разметок. Все размечаемые во всех текстах сущности получают

<sup>6</sup> Один из подступов к этой проблеме — электронная версия указателя персоналий и географических названий из 91-го тома Полного собрания сочинений Толстого: http://index.tolstoy.ru/ (accessed April 14, 2017).

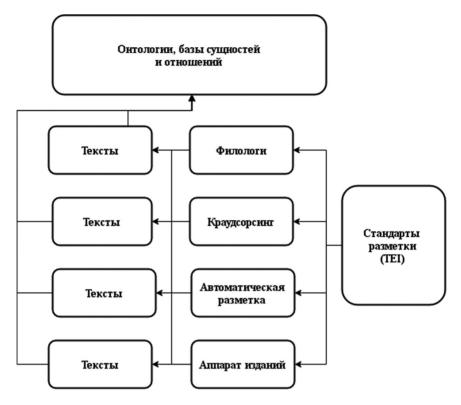


Рис. 2 Архитектура семантической сети культуры

уникальный идентификатор; свойства сущностей и отношения между сущностями, как и в классической семантической сети Бернерса-Ли, хранятся в универсально доступных базах (основой базы идентификаторов ресурсов и онтологий может стать Википедия).

Насколько утопичен этот проект? Нам кажется, движение к связному семантическому представлению в каком-то виде (наверняка не в точности так, как мы описали) неизбежно. Как уже сказано, несмотря на (быть может, временную) неудачу глобального проекта семантической сети, все больше областей информации оказываются охвачены ее элементами и идеологией «связанных данных». Культурное наследие — одна из таких областей. Несколько важных тенденции в филологии, лингвистике, и компьютерном анализе данных последнего десятилетия можно считать предпосылками для появления семантических изданий и семантической сети культуры.

Прежде всего, в последние десятилетия как лингвисты, так и литературоведы активно используют так называемые корпуса, т.е. размеченные электронные собрания текстов с надстроенным поисковым механизмом. Несмотря на претензии филологов к устройству этих ресурсов<sup>7</sup>, они связывают данные воедино, дают широкую панораму словоупотребления, в том числе и в культурно значимых текстах. Некоторые корпуса уже включают семантическую разметку, другими словами, позволяют искать, например, всех животных, растения, имена родства, еду и напитки, а также разные другие сущности<sup>8</sup>. Семантическое издание в предлагаемом нами смысле можно рассматривать как разновидность корпуса с некоторыми существенными добавлениями и надстройками, важнейшие из которых — связанность со всеми остальными корпусами посредством унифицированной разметки и общей базы сущностей. Семантическую сеть культуры можно определить как корпус всех корпусов.

Движение кунификации разметки — другая важная составляющая семантического Zeitgeist'а. Платформой для семантических изданий, по всей видимости, станет TEI (Text Encoding Initiative): это сообщество специалистов и разработанные этим сообществом правила кодирования текстов на основе формата XML. Сайт TEI9 содержит тысячи страниц описания принципов и конкретных примеров разметки, призванных унифицировать представление данных и максимально упростить взаимную конвертацию форматов. TEI (созданный еще в 1980-е) становится стандартом электронного представления текстов, а вокруг этого стандарта формируется набор полезных для исследователей и электронных издателей инструментов, таких, как облегчающие кодирование расширения

<sup>7</sup> См. А.С. Бодрова, И.А. Пильщиков, "Проблемы корпусного подхода к задачам авторской лексикографии", Л.Л. Шестакова (отв. ред.), Авторская лексикография и история слов: К 50-летию выхода в свет «Словаря языка Пушкина» (М.: Азбуковник, 2013), 59–61; Б.В. Орехов, "О перспективах филологического корпуса", Труды Отделения историко-филологических наук РАН. 2015 (М.: Наука, 2016), 147–155.

<sup>8</sup> Категории семантической разметки, доступные в Национальном корпусе русского языка, перечислены на этой странице: http://ruscorpora.ru/reqsem.html (accessed April 14, 2017).

<sup>9</sup> http://www.tei-c.org/ (TEI Consortium, 2007-...).

для программ-редакторов текста, веб-сервисы, реализующие автоматическое преобразование между разными форматами $^{10}$ .

Семантическая сеть предполагает наличие постоянно универсальной базы сущностей, поддерживаемой сущностей, и онтологий извлекаемых из текстов. Википедия (вернее ее представление в виде открытых машиночитаемых данных, DBpedia) — это наиболее логичный и общепризнанный претендент на роль такой базы. Наконец, в самые последние годы на наших глазах происходит очередная революция в области автоматического анализа данных и машинного обучения. Методы глубинного обучения и нейронных сетей позволяют автоматически размечать все более и более сложные сущности, обнаружение которых еще несколько лет назад требовало бы квалифицированной ручной работы историков и филологов.

Мы верим, что эти тенденции со временем приведут к появлению семантических изданий, которые затем сольются в единую семантическую сеть. Верим и призываем миллионы текстов поскорее обняться. А пока этого не произошло, с ранними прототипами инструментов семантической сети культуры можно познакомиться на примере существующих проектов, основанных на близких идеях и технологиях.

# Примеры семантических изданий и инструментов

У текстов, и у отдельных элементов текста, есть временное измерение: датировки, упоминаемые даты, внутренняя хронология и т.д. Разметка этих данных свяжет одновременные события в разных текстах и приведет к появлению универсальной хронологии культуры. Прообразом такого инструмента в применении к русской дневниковой прозе стал электронный корпус личных дневников «Прожито»<sup>11</sup>. Этот ресурс выполняет главную функцию семантического издания — позволяет текстам взаимодействовать друг с другом. Хронологическая разметка дневниковых записей

<sup>10</sup> Пожалуй, двумя главными препятствиями на пути неизбежного рождения всеобщего семантического издания в будущем стоят разнообразие форматов кодирования текста и авторское право. ТЕІ в той или иной степени дает оружие для борьбы с первой из этих сложностей.

<sup>11</sup> http://prozhito.org/ (М. Мельниченко [и др.], 2015-...).

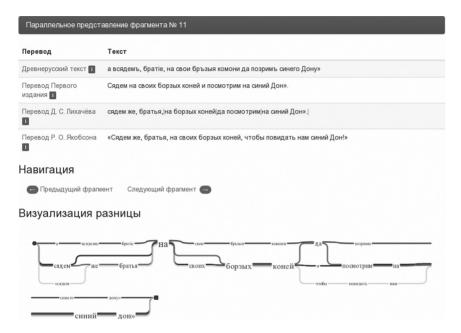


Рис. 3 Визуализация разницы переводов «Слова о полку Игореве»

сводит вместе все записи, сделанные в один и тот же день, и, тем самым, воссоздает экзистенциальную панораму этого дня, дает возможность «прожить» его, погрузившись в опыт и переживания современников. Добавление хронологической разметки материалов прессы и кинохроники сделает подобные проекты еще более информативными и многомерными.

Пример более традиционной филологической разметки — параллельные тексты и переводные эквиваленты. На основе такой разметки построен «Параллельный корпус переводов "Слова о полку Игореве"»<sup>12</sup>. «Слово», как известно, текст со сложной судьбой: когда рукопись была утеряна в московском пожаре 1812 года, памятник остался без материального носителя, то есть без надежного источника. Во-первых, это способствовало живучести проблемы подлинности. Во-вторых, породило множественность гипотетических конъектур и интерпретаций конкретных мест в тексте. В итоге «Слово» стало многослойным культурным артефактом, без общепринятого эталонного облика, кроме того,

<sup>12</sup> http://nevmenandr.net/slovo (Б. Орехов, 2007-...).

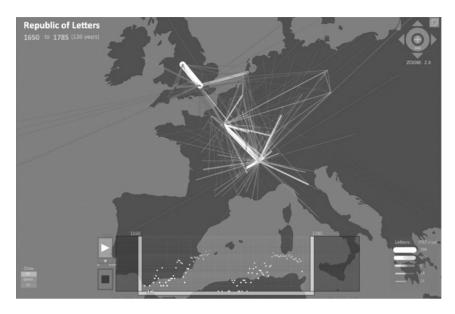
требующим дополнительной репрезентации — в виде перевода на современный русский язык. Таких переводов существуют сотни, а различия между ними исторически обусловлены, а значит, культурно значимы, и взывают к исследовательскому вниманию. Корпус переводов «Слова» представляет их в «выровненном», параллельном, виде: каждому месту в каждом переводе сопоставлен эквивалентный ему отрезок текста в другом переводе, тексты оказываются связаны не номинально (по названию), а непосредственно. Сопоставление выполнено с помощью специального вида разметки, фиксирующего т.н. «выравнивание» (см. рис. 3). Всё это помогает исследователю сравнивать переводы, уточнять их разницу, искать этой разнице историко-культурное объяснение.

Еще один проект такого рода: корпусное параллельное французской представление поэзии В русских переводах первой трети XIX века<sup>13</sup>, позволяющее выявлять особенности межкультурного взаимодействия в стихах: в корпусе видно, что и как поэты передают на своем языке, что теряют (намеренно или под действием условий языка или поэтической традиции), а что добавляют к исходному тексту.

В Разметка персоналий массивах текстов позволяет восстанавливать социальные и интеллектуальные сети прошлого и исследовать их динамику. Одним из первых проектов такого рода стал сайт «Mapping the Republic of Letters» 14 представляет в виде электронного ресурса размеченное эпистолярное наследие деятелей эпохи Просвещения. Разметка включает информацию об адресате, месте и времени отправления и получения письма. Таким образом, создатели сайта смогли построить впечатляющую анимированную карту, показывающую движение корреспонденции (и, фактически, стоящих за ней идей) в Западном мире в XVII—XIX вв. Как показано на сайте в разделе частных исследовательских применений получившегося ресурса, уже в существующем виде разметка позволяет проверить некоторые гипотезы и устойчивые представления, например, касающиеся космополитизма тех или иных авторов. В частности, можно сравнить «сеть» по переписке нескольких деятелей Просвещения.

<sup>13</sup> http://nevmenandr.net/fr/ (А. Бодрова, Б. Орехов, 2011-...).

<sup>14</sup> http://republicofletters.stanford.edu/ (Stanford University, 2013-...).



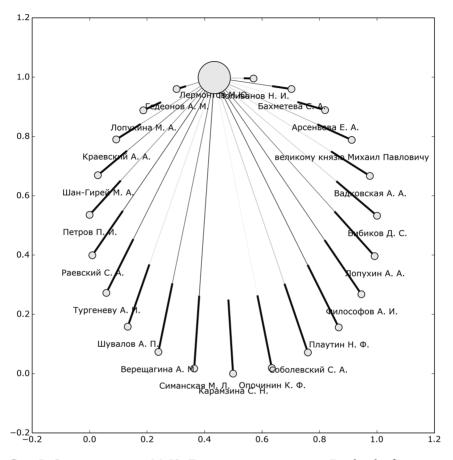
**Рис. 4** Карта корреспонденции Дж. Локка и Вольтера из проекта «Mapping the Republic of Letters»

Подобный же подход — разметка переписки русских классиков XIX века и попытка реконструкции социальных связей литературного сообщества — представлен в проекте «Facebook прошлого» 15 (см. рис. 5).

Социальные сети можно восстанавливать не только по эпистолярных данным, но и при помощи статистического анализа совместной встречаемости размеченных персоналий. Пример такого подхода — социальные карты деятелей культуры на основе разметки исторических изданий энциклопедии Британника и Википедии в проекте Эволюция Знания (Knowledge Evolution Project)<sup>16</sup> (см. рис. 6).

<sup>15</sup> http://web-corpora.net/Facebook\_of\_the\_Past/facebook.html (М. Гронас, Б. Орехов [и др.], 2015-...). На нынешнем этапе трудность в том, что собрания сочинений ориентированы на изолированные связи писателя со своими корреспондентами, то есть мы не можем по опубликованным там материалам восстановить «горизонтальные» отношения тех, кто общался с авторами, между собой. Поэтому получается неполная схема сети, отраженная на иллюстрации ниже. В будущем, когда будут оцифрованы и размечены пропущенные в академических собраниях тексты, эти недостающие социальные связи будет возможно восстановить и дополнить имеющуюся картину.

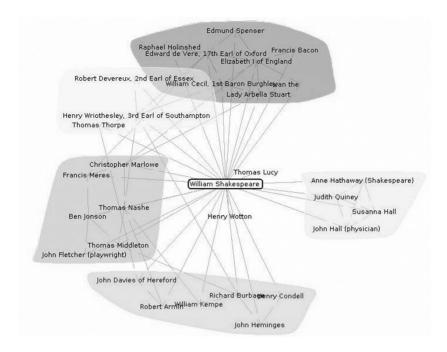
<sup>16</sup> http://knowevo.cs.uml.edu/ (Mikhail Gronas, Anna Rumshisky, Alexey Romanov, Yen-Fu Luo 2013-...).



**Рис. 5** Социальная сеть М. Ю. Лермонтова из проекта «Facebook of the Past»

Учитывая разметку хронологических отношения между упоминаниями исторических лиц и характер цитирования можно выстроить карту интеллектуальных связей (см. рис. 7).

В этом же проекте сделана попытка семантического «ресайклинга» — т.е. превращение в машиночитаемую разметку данных накопленных за 200 лет в каталоге Библиотеки Конгресса США. Каталог содержит библиотечные шифры (легко сводимые к тематической и хронологической разметке) и ежегодные реестры приобретения книг. Анализ этих данных позволяет визуализировать глобальные тенденции в развитии знания: рост и угасание, появление и исчезновение предметных областей, отношения между



**Рис. 6** Социальная карта Шекспира по энциклопедиям XIX–XXI вв.

областями. На иллюстрациях ниже представлены карты областей (галактики) знания в Библиотеке Конгресса по данным на 1881, 1912, 1947 годы (см. рис. 8, 9, 10). Каждый кружок — предметная область или дисциплина. Размер кружка соответствует пропорциональному размеру дисциплины (процент от всех книг в коллекции библиотеки). Близость областей на карте передает степень взаимной цитации между областями, т.е. интердисциплинарные связи. Карты иллюстрируют исторический процесс интердисциплинарного перемешивания в конце XIX и первой половине XX века. В 1881 году области наук автономны, а к 1947 году взаимосвязаны и переплетены 17.

Те же данные использованы в «геологической» карте знаний для иллюстрации разнице в скорости накопления знаний по предметным областям (рис. 11, 12)<sup>18</sup>.

<sup>17</sup> См. интерактивную анимированную карту: http://knowevo.cs.uml.edu/demos/congress-lib/v3/

<sup>18</sup> Анимированная карта: http://knowevo.cs.uml.edu/demos/congress-lib/surface/

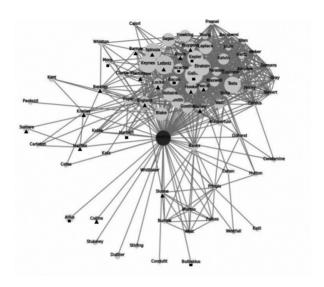


Рис. 7 Интеллектуальная сеть Исаака Ньютона. Квадратом помечены ученые, которые повлияли на Ньютона, треугольником — современники Ньютона, остальные узлы его последователи (Knowledge Evolution Project)

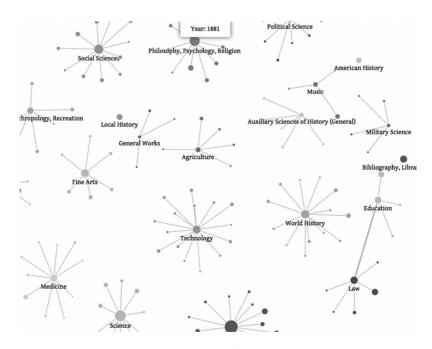


Рис. 8 Галактика знаний по данным Библиотеки Конгресса: 1881

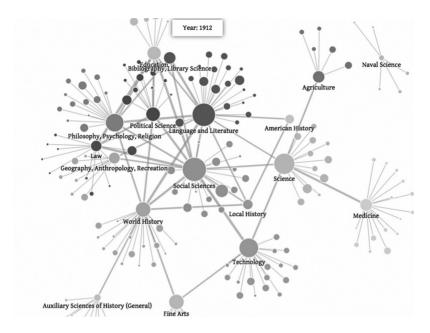


Рис. 9. Галактика знаний по данным Библиотеки Конгресса: 1912

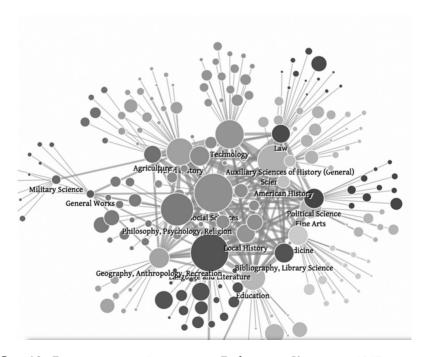


Рис. 10. Галактика знаний по данным Библиотеки Конгресса: 1947

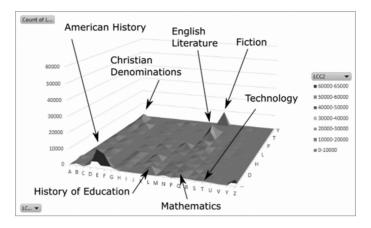


Рис. 11 Рельеф знаний по данным Библиотеки Конгресса, 1900

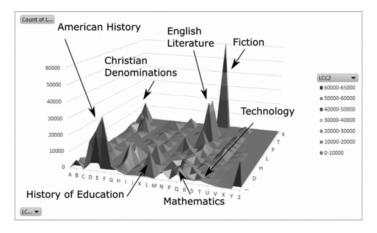
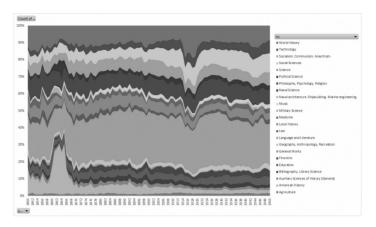


Рис. 12 Рельеф знаний по данным Библиотеки Конгресса, 1950

Еще одна проекция тех же данных, «пульс знания», позволяет соотносить изменения в удельным весе дисциплин с историческими процессами (рис. 13)<sup>19</sup>.

Обратим внимание читателя на расширение полосы, представляющей Американскую историю (вторая снизу), в 1861–1866 гг., и расширение полосы, представляющей всемирную историю (первая сверху), в 1914–1919 гг. Очевидно, что рост интереса к

<sup>19</sup> Интерактивная визуализация: http://knowevo.cs.uml.edu/demos/congress-lib/rivers/



**Рис. 13.** Пульс знания: распределение книг в Библиотеке Конгресса по предметным областям (1850–1950)

истории в обоих случаях вызван войнами: гражданской войной в США и первой мировой.

К описываемому здесь образу семантического издания стремится и инициатива Tolstoy Digital, подразумевающая превращение толстовского текста в данные. На текущем этапе уже осуществлен перевод 90-томного издания в формат  $TEI^{20}$ . Ср. также концепцию представления «Калевалы» в виде семантической сети<sup>21</sup>, и семантическое представление социальной сети раннего Нового времени в Англии в проекте «Six Degrees of Francis Bacon»<sup>22</sup>.

#### **NO PASARAN**

Приведенные примеры могут показаться яркими (особенно когда они снабжены визуализациями), но не слишком интересными с научной точки зрения. Скептик должен усомниться: зачем было городить огород, сливать воедино все написанное, размечать сущности вручную и автоматически, если результаты — карты, социальные

<sup>20</sup> http://digital.tolstoy.ru/tolstoy\_search/data/ (accessed April 15, 2017).

<sup>21</sup> Eero Hyvönen, Tuomas Palonen, Joeli Takala, "Narrative Semantic Web — Case National Finnish Epic Kalevala", *Poster Papers, the 7th Extended Semantic Web Conference, Heraklion, Greece*, June, 2010, URL: http://seco.cs.aalto.fi/publications/2010/hyvonenet-al-narrative-sw-2010.pdf (accessed April 15, 2017).

<sup>22</sup> http://www.sixdegreesoffrancisbacon.com/ (Carnegie Mellon University, 2015-...).

сети авторов и персонажей, хронологии, списки параллельных мест — по большому счету тривиальны, поверхностны и не слишком приближают нас к пониманию культурных и исторических феноменов. Мы отчасти согласимся со скептиком: действительно многие исследования такого рода пока разочаровывают. Но напомним, что этот подход к культуре появился совсем недавно, мы наблюдаем пока младенчество изучения сети культуры. В примерах, которые мы привели, задействованы самые простые, лежащие на поверхности данные (даты, имена, названия дисциплин). Тем не менее, многие из полученных результатов далеко не тривиальны (и не могли бы быть получены в досемантическую эпоху). А уже очень скоро будет задействована глубинная семантическая разметка добываемая, например, нейронными сетями. Тогда мы сможем автоматически отвечать на самые разные действительно сложные и интересные вопросы: О чем думал, чем занимался, что написал автор в тот или иной конкретный день или час? О чем чаще всего думают герои книг, когда смотрят на небо? Как, при каких обстоятельствах, отчего смеялись, плакали, падали в обморок в XVIII, XIX, XX веке? Почему изменилось отношение к публичной казни в эпоху Просвещения? Как менялись механизмы внимания и памяти в истории нашей цивилизации? И этот последний вопрос подводит нас к самому главному:

Компьютер и Интернет — безусловное зло. Эти технологии делают нас невнимательными и забывчивыми. Поскольку спрятаться от них совершенно невозможно, единственный способ борьбы — это тушить пожар пожаром. Семантическая сеть культуры и новые семантические инструменты, которые уже возникают и еще возникнут — это способ вернуть культуре связность и единство, одно из орудий (отобранное у врага!) в партизанской войне против электронного забвения и рассеяния.