Формализация фактоподобных высказываний в конкретно-исторических исследованиях

© Н.А.Маркова
Институт проблем информатики РАН,
Москва
MarkovaNatAlex@gmail.com

Аннотация

На основе анализа специфики конкретно-исторических исследований разработана модель представления фактоподобных высказываний, включающих не только точные утверждения, но и неполные сведения, результаты их аналитико-синтетической обработки, вопросы и гипотезы. Представление высказываний в виде метаданных является основой поддерживающей информационной технологии.

1 Введение

Массовые электронные публикации исторических источников и исследований открывают широчайшие возможности для работ по изучению конкретно-исторических вопросов. Для того чтобы ввести в научный оборот публикуемые материалы, требуется провести их фактографическое индексирование, оснастить метаданными, представляющими содержащие в них сведения – факты – в удобном для использования виде. Эта задача жизненно важна не только для рукописных и старопечатных документов, но и для поддающихся переводу в полнотекстовый вид нарративных источников. В той или иной степени ее решает каждый исследователь, изучая источник. Не дожидаясь пока библиографы и архивисты осуществят фактографическое индексирование источников, эту работу уже выполняют виртуальные сообщества исследователей, как профессиональных, так и любителей. Подавляющее большинство площадок для обмена фактографической информацией представляет собой бессистемный обмен текстовыми репликами на форумах.

Однако существуют примеры и хорошо продуманных информационных технологий в этой области. Фактографическое индексирование, выполняемое виртуальным сообществом в рамках крупнейшего международного проекта FamilySearch [5], насчитывает более миллиарда записей, в подготовке которых участвуют сотни тысяч волонтеров. В проекте фигурирует ограниченный круг хорошо структурированных источников (в основном, регистрационных), и выходом его служит ограниченная номенклатура фактов – основные даты биографий лиц и их родственные связи.

Чрезвычайно интересные результаты были получены в рамках проводимого в Петрозаводском университете комплекса работ по формализации информации, содержащейся в коллекциях текстов исторических документов, и построения информационной системы для упорядочивания и анализа накопленных знаний в рамках работы сетевого сообщества [2]. Модель предполагает глубокую и множественную разметку исходных документов. Её сфера применения в настоящее время ограничена сообществами исследователей рукописных средневековых текстов.

Ряд особенностей конкретно-исторических исследований не позволяет применить унифицированные готовые решения, опирающиеся на представлении о факте, как об утверждении, что, в частности, неявно предполагает семантический Web. Далеко не все факты, излагаемые в исторических источниках, в основанных на них исследованиях, а также в справочниках и энциклопедиях, соответствуют объективной истине. Документы нередко содержат предположения, гипотезы, частичное знание об интересующем предмете. При этом приближение к истине возможно за счет анализа противоречий, интеграции данных, извлекаемых из различных источников. Метаданные нередко используются для представления фактографической информации (например, в проекте dbPedia), однако в них не учитываются чрезвычайно важные для конкретно-исторических исследований особенности темпоральность и неточность.

Расширим понятие «факт», включив в него неточные и неполные сведения, результаты их аналитико-синтетической обработки, вопросы и гипотезы. Предложим общую форму для фиксации такого рода сведений в виде фактоподобных высказываний (ФПВ), представимых метаданными. Модель объединяет

данные фактографического индексирования исторических источников и их аналитико-синтетической обработки. Основные положения предлагаемого формализма будем выражать в терминах ER-модели, на концептуальном уровне совпадающих с категориями аппарата онтологий, повсеместно применяемых в настоящее время для формального представления фактических знаний.

Наши построения будут основаны на анализе специфики конкретно-исторических исследований. Их целью является создание основы для построения эффективной информационной технологии поддержки работы исследователей.

2 Специфика конкретно-исторических исследований

В рамках конкретно-исторических исследований изучаются определенные объекты, сведения о которых частично формализуемы. Предполагается, что в данной сфере имеется специальный (выбранный исследователем) понятийный аппарат, и предметом исследования является вполне определенный набор свойств объектов, часть из которых может быть определена математически множеством допустимых значений, а часть характеризуется нарративами или образами.

Мы рассматриваем исследования, опирающиеся на изучение документальных источников. Диапазон такого рода работ постоянно расширяется за счет того, что существенная часть документов получает электронные копии, к которым обеспечивается сетевой доступ.

Изучая источник, исследователь сохраняет метаданные: адресные ссылки, выдержки, выписки, а также, по возможности, некоторую формализованную в соответствии с задачами конкретного исследования форму извлеченного знания. В соответствии с классификацией, данной в работе [1], метаданные делятся на автономные и встроенные. В терминах традиционной бумажной технологии первые – соответствуют записям, сохраняемым в виде отдельных карточек или в рабочей тетради. Вторые – результат разметки документа-источника – очерчивания, закладок, заметок на полях, использования разноцветных маркеров или стикеров.

При переходе к современной информационной технологии эффективность работы исследователя будет тем выше, чем более систематизировано удастся представить эти метаданные.

Перечислим основные особенности изучаемых объектов, которые следует учитывать при создании информационной технологии, обслуживающей конкретно-исторические исследования.

- Имена объектов вариативны (объект может иметь несколько имен) и неоднозначны (различные объекты могут иметь совпадающие имена).
- Период существования объекта, а также периоды, в котором значение некоторого свойства его постоянно, представляют ограниченные временные интервалы.
- Номенклатура изучаемых свойств объектов специфична для определенного класса объектов и зависит от конкретного исследования. Причем наличие определенного свойства и его возможные значения, а также допустимые сочетания значений этого и других свойств объекта зависит от временного интервала даже в рамках одного исследования.
- При определении свойств объекта возможны искажения вследствие дефектов в содержании источников, в процессах их распознавания, интерпретации, интеграции.

С точки зрения процесса исследований к поддерживающей его информационной технологии целесообразно предъявить следующие требования.

- Необходимо фиксировать не только четко установленные факты, но и ФПВ, включающие предположения, неточные значения свойств, исследовательские вопросы.
- Каждое сформулированное высказывание должно быть соотнесено с источником или с цепочкой вывода, обобщающей другие ФПВ.
- Необходимо обеспечить информационную навигацию по связям между объектами, в том числе, для электронных документов – межтекстовые связи; многоаспектный поиск; возможности статистической обработки.
- Необходимо отслеживать процессы накопления данных, выявления дефектов, выдвижения/ опровержения гипотез по исследуемым источникам, времени, исполнителям, аргументации.

Концептуальной основой для создания информационной технологии, удовлетворяющей перечисленным требованиям, является предлагаемая формальная модель представления ФПВ.

3 Модель фактоподобных высказываний

Принципиальная проблема, которую необходимо решить при разработке модели ФПВ, состоит в выборе рационального уровня формализации. Малоэффективны как совсем неформальное текстовое представление (нарратив), так и попытка максимальной формализации. Сформулируем три положения, опора на которые позволит выбрать оптимальный уровень формализации.

1) Модель строится по ER-принципу, с определенными наборами объектов, атрибутов, отношений.

2) Для каждой сферы исследования выбираются свои наборы объектов и свойств, возможно, уточняемые для конкретного проекта.

3) Формализуются не все возможные свойства, а те, которые отражают поддающиеся типизации аспекты, возможные значения которых задаются диапазоном чисел, дат; словарным перечнем. Все, что не укладывается в эти рамки (а также малозначимые в рамках конкретного проекта сведения), представляется нарративным текстом.

Модель включает три группы элементов. Основные элементы модели – компоненты ER-модели изучаемого исторического процесса: объекты, атрибуты, отношения – фиксируются базовыми высказываниями. Высказывания-связки соотносят базовые высказывания с источниками и между собой. Наконец, информацию, включающую высказывания-ограничения, а также данные, относящиеся к процессу исследования, отнесем к служебным высказываниям. Рассмотрим эти группы высказываний подробнее.

3.1 Базовые высказывания

Множество объектов исследования ($O = \cup O_{class}$) включает объекты определенных классов. Для каждого класса объектов устанавливается набор свойств. Литеральные свойства – атрибуты – сопоставляют объекту некоторое значение из определенного множества (чисел, дат, номинальных шкал, текстов). Объектные свойства – отношения – сопоставляют объекту другой объект и литеральное значение, которое можно воспринимать, как метку на графе связей между объектами. Такая конструкция, вместо используемого в OWL строгого разделения на категории свойств, не предполагающего литеральных значений у объектных свойств, позволяет не вводить дополнительных объектов (*Отношение между Петровым и Гимназией*), а, оставаясь в рамках объектов исследования (*Петров*, *Гимназия*) специфицировать значение связи (*Должность = Инспектор*). Для конкретно-исторических задач такое представление существенно нагляднее.

Далеко не все высказывания, содержащиеся в историческом источнике, можно формализовать. Но даже для формализуемых высказываний, суждение о значении свойства объекта может быть сформулировано не только как равенство некой константе, но и как различные варианты неравенства, а также принадлежности (не принадлежности) некоторому набору констант. Оператор ФПВ, соотносящий значения свойства объекта с константой/списком констант (•), определим следующим образом:

$$\bullet \in \{=, \approx, \neq, <, >, \in, \notin\}$$

Наиболее важная особенность предлагаемой модели – включение в ФПВ временного интервала, в рамках которого оно предполагается справедливым. Такая конструкция предоставляет значительно более удобный базис для аналитико-синтетической работы в области конкретно-исторических исследований, чем фиксация отдельных событий. Действительно, подавляющее большинство событий, касающихся объекта, имеют свою пару – они фиксируют начало и конец периода, в котором некоторое свойство объекта имело некоторое значение. Даже для такого свойства, как титул, возможны события присвоение, лишение, восстановление, определяющие соответствующие временные интервалы. В любом случае, время ФПВ, касающегося некоторого свойства объекта, ограничено временем существования (жизни) объекта.

Будем определять периоды ($dt \in DT$), как

$$dt = (start, finish),$$

где start и finish – это либо даты (с некоторой степенью точности), либо оценки ограничений, налагаемые на эти даты. Подробно форма представления временных интервалов разной степени определенности в виде строки метаданных рассмотрена в [3].

Предложенный подход совсем не противоречит возможности в рамках конкретного исследования определить специальный класс объектов – события определенного рода (например, *Конференции*).

Рассмотрим основные виды базовых ФПВ и определим содержание метаданных, их фиксирующих.

1) Дефиниция – высказывание, определяющее существование объекта определенного класса в определенный период времени:

$$\forall t \in dt \ (o_d(t) \in O_{class}).$$

d = (nomen, class, dt) – метаданные, фиксирующие дефиницию.

Здесь nomen – имя объекта – неформальная и, возможно, неуникальная текстовая константа, служащая для удобства восприятия исследователем. Каждое ФПВ имеет свою уникальную идентификацию, которую для простоты описания мы опускаем. При адресации ФПВ будем использовать его обозначение (например, d).

2) Атрибут – высказывание, определяющее значение определенного литерального свойства объекта в определенный период времени:

$$\forall t \in dt \ (a_{aclass}(o_d, t) \bullet avalue).$$

a = (d, aclass, avalue, •, dt) – метаданные, фиксирующие атрибут.

3) Отношение – высказывание, определяющее связь объекта с другим объектом, а также

литеральное значение, сопоставляемое этой связи в определенный период времени:

$$\forall t \in dt \ (r_{rclass}(o_{dp}, o_{dq}, t) \bullet rvalue).$$

r = (d_p, d_q, rclass, rvalue, •, dt) – метаданные, фиксирующие отношение.

Во всех конкретно-исторических исследованиях рассматриваются классы *Лицо* и *Документальный* объект (*Д-объект)*. В большинстве случаев интерес представляют *Географические* и *Социальные* объекты. В специальных исследованиях классами изучаемых объектов являются *Архитектурные, Природные, Математические и пр.* объекты.

Наиболее общими для самых разных областей исследований являются свойства *Д-объектов*, под которыми мы понимаем не только документы, но и их совокупности, и их компоненты (от архивов, библиотек, интернет-порталов до абзацев текста). Атрибуты и связи документов хорошо специфицируют библиографические и археографические стандарты. В рамках современных стандартов IFLA (например, [5]) рассматриваются связи между документами, представляющие интерес для конкретно исторических исследований. К ним относятся: структурная (входит, следует за), деривативная (версии, переработки, переводы), дескриптивная (критика, комментарии, аннотации, рефераты) связи.

Атрибуты и связи *Лиц*, в основном, специфичны для сферы исследований. Универсальны атрибуты *пол* и связь с гео-объектами *местопребывание* (которое, например, в момент рождения – место рождения). Достаточно часто рассматриваются родственные связи, должностные отношения, отношения учитель-ученик. Связи *Лиц* и *Д-объектов* фиксируют сведения об авторстве, адресатах и упоминаниях.

3.2 Высказывания-связки

Утверждение о том, что некоторое ФПВ получено в результате интерпретации (⇒) определенного источника также является определенным высказыванием. Источник при этом адресуется дефиницией соответствующего *Д-объекта*. Сопоставляя ФПВ, исследователь конструирует новые выражения с помощью логических или темпоральных связок. Как интерпретация, так и логические операции над ФПВ не являются в полной мере формальными действиями. В рассуждениях исследователя есть доля интуиции. Однако степень уверенности в своих умозаключениях вполне оцениваема. Поэтому каждому ФПВ-связке будем сопоставлять оценку уверенности в фиксируемой им формулировке. Такую оценку рационально выражать в шкале нечеткой логики от 0 – FALSE до 1 – TRUE.

Пятно на рукописи, неразборчивый почерк, неизвестные сокращения – причины того, что исследователь неуверен в результатах интерпретации. Но и при уверенности в толковании источника, исследователь может быть не согласен со смыслом интерпретированного высказывания. В этом случае он должен зафиксировать противоречие между данными источника и более надежными сведениями, что послужит обоснованием для высказывания, фиксирующего ложность сведений источника. Пример цепочки такого рода размышлений, фиксируемых средствами ФПВ, будет приведен в следующем разделе.

Определим множество ФПВ (представленных метаданными) – F, как объединение вышеперечисленных видов ФПВ и специальных ФПВ-связок – L, определяемых рекурсивно.

$$F = D \cup A \cup R \cup L,$$

где $D = \{d\}$, $A=\{a\}$, $R=\{r\}$, $L = \{l\}$

$l = (fp, fq, \blacklozenge, estim)$

$fp \in F$, $fq \in F$,

$estim \in [0..1]$, (0 –TRUE, 1 – FALSE)

$\blacklozenge \in \{\Rightarrow\} \cup Logical \cup Temporal$

Logical = {AND, OR, XOR, …}

Temporal = {BEFORE, AFTER, SAMETIME, INTERSECT}

⇒ – интерпретация.

3.3 Служебные высказывания

Каждой сфере исследования соответствует свой набор классов объектов, их свойств, зависимостей между значениями свойств. Часть из этих ограничений легко формализуема. Например, спецификация перечней классов объектов и классов свойств, в зависимости от классов объектов; списки возможных значений свойств. Несколько сложнее, но все же возможно формализовать ограничения на возможные сочетания значений свойств, а также на временные характеристики. Примерами такого рода ограничений являются накладываемые биологическими законами разности в возрасте родителей и детей, или формулируемые конкретным социальным устройством регламент продвижения по службе.

Важнейшим служебным высказыванием является перечень классов (например, class ∈ {*Лицо*, *Д-объект*, *Гео-объект*, *Соц-объект*}). Ограничения на атрибуты формулируются указанием области определения и области значений (domain и range). Например, domain(*Пол*)= *Лицо*; range(*Пол*)={м, ж, ?}. Для отношений область определения задается парой, например, domain(*Родство*)= (*Лицо*, *Лицо*).

Для фиксации ограничений на возраст детей может потребоваться формализация высказывания

$\forall$ d_0, d ((d_0, d, *Родство*, *Родитель*) $\rightarrow$
(d.start – d_0.start > 10) AND
(d.start – d_0.start < 90))

Должны ли фиксироваться подобные ограничения в виде метаданных, интерпретируемых некоторым унифицированным инструментом, или они представляют специализированные процедуры контроля (своего рода сложных алгоритмических высказываний) – зависит от конкретных обстоятельств. Во многих случаях контроль ограничения вообще может быть выполнен только вручную. В целом, полезно хотя бы в неформальном, текстовом виде фиксировать ограничения, как своего рода памятку для исследователя (нарративное высказывание).

В соответствии с выдвинутыми требованиями необходимо отслеживать процесс накопления данных и их аналитико-синтетической обработки. Для этого целесообразно применить типовой прием, используемый, в частности, в wiki-технологии. Каждая запись ФПВ сопровождается временной меткой и указанием автора. Вместо изменения записи производится формирование ее новой версии.

4 Пример рассуждений, фиксируемых ФПВ

Рассмотрим пример интерпретации источника, выявления противоречия, формулировки новых ФПВ. Источник – книга, посвященная 100-летию Первой московской гимназии [6].

d_0 = («Столетие 1-й гимназии», *Д-объект*, 1903)

d_1 = («1-я гимназия», *Соц-объект*, 1804-1904..)

В источнике содержатся, в частности, списки выпускников по годам выпуска, а также списки печатных работ, авторами которых являются выпускники гимназии. Два однофамильца – Алексей М. и Александр М. окончили гимназию соответственно в 1896 и в 1888 годах:

d_2 = («Алексей М.», *Лицо*, 1874..1878-1903..)

r_1 = (d_2, d_1, *Ученик*, ..-1896)

d_3 = («Александр М.», *Лицо*, 1866..1870-1903..)

r_2 = (d_3, d_1, *Ученик*, ..-1888)

Оценка времени жизни дана, исходя из ограничения на возраст учеников.

В источнике допущена ошибка. В комментарии, относящимся к Александру М., сказано «Известный этнограф, исследователь былин сев. края». При этом работ у Александра М. не отмечено, а вот у Алексея М. отмечено несколько работ, посвященных северным былинам.

a_1 = (d_3, «исслед. былин сев. края», 1888..-)[1]

l_1 = ($d_{0\ c.295}$, a_1, $\Rightarrow$, 1)

a_2 = (d_2, «автор Беломорские былины», 1901)

l_2 = ($d_{0\ c.295}$, a_1, $\Rightarrow$, 1)

Итак, ФПВ a_1 и a_2 противоречивы:

l_3 = (a_1, a_2, *AND*, 0.01)

Формальную возможность того, что и Александр М. был «известным этнографом», но не публиковал своих исследований, мы оценили в 1 процент.

l_4 = (l_3, a_2, $\Rightarrow$, 0.01)

l_5 = (l_4, r_1, *AND*, 0.99)

Теперь мы можем сформулировать новое высказывание, корректирующее ошибочное a_1:

a_3 = (d_2, «исслед. былин сев. края», 1896..-)

l_6 = (l_4, a_3, $\Rightarrow$, 0.99)

Строго говоря, приведенная цепочка рассуждений, равно как и операция интерпретации источника не являются формальными. Однако возможность формализованной фиксации результатов мыслительных операций существенно дисциплинирует исследователя, а также позволяет осуществиться научной коммуникации, что служит залогом взаимного контроля и способствует повторному использованию данных исследования.

5 Заключение

В рамках данной работы модель ФПВ представлена концептуально. При ее использовании в конкретной информационной технологии она должна быть выражена в терминах соответствующего аппарата, в качестве которого могут выступать как современные языки онтологий, так и инструменты баз данных.

Опора на языки онтологий позволит организовать обмен информацией с другими информационными системами. В частности, это позволит импортировать конечные (или хотя бы стабилизированные) данные исследования в качестве фактографического индекса в библиографическую/археографическую информационную систему.

Технология баз данных обеспечит эффективность накопления и аналитико-синтетической обработки ФПВ. Однако наилучшего результата, как показала практика разработки и эксплуатации инструментального комплекса Фактограф [4], можно добиться, сочетая автономные метаданные, хранимые в базе данных, и встроенные, размечающие документ-источник. При этом предполагается, что исследователь имеет свою копию источника, которую он может «чиркать» разметкой. Взаимные связи между ФПВ, хранимыми в базе данных, и фрагментами текста источника достигаются средствами гиперссылок. В документах-источниках границы фрагментов,

[1] Для краткости мы опускаем • = «=», класс атрибута – *Упоминание*, а также определения Д-объектов – страниц, входящих в книгу-источник.

связанных с высказываниями, хранимыми в базе, определяются либо явно (для xml и html форматов), либо закладками, применимыми не только в офисных документах, но и в документах форматов pdf и djvu. В свою очередь, гиперссылками на форму, представляющую конкретный объект в базе данных, целесообразно оснастить текст источников в точках его упоминания. В случае не редактируемых документов (pdf и djvu) такую ссылку можно поместить в комментарий.

Вычленяя ФПВ из источника, мы обеспечиваем удобство его контроля, анализа, интеграции, но в то же время, теряем контекст, который может быть чрезвычайно полезен для создания целостной картины. С другой стороны, возможность получения оперативной справки по ходу чтения источника, касающейся его текущего участка, способствует пониманию текста. Сравнение обладающего внутренним единством линейного текста со структурной картиной связанных объектов, в нем упоминаемым, дает возможность как уточнить идентификацию объектов, сформулировать новые ФПВ, так и глубже понять подтекст, неподдающийся формализации.

Повторное обращение к источнику (адресное, и поэтому эффективное), равно как и повторное использование выявленных сведений, чрезвычайно полезно уже индивидуальному исследователю. Тем важнее эти возможности для организации информационного обмена в сообществах, изучающих историю. Предложенный в работе метод формализации данных может служить основой для создания информационной технологии, существенно повышающих эффективность работы коллектива исследователей.

Литература

[1] Когаловский М.Р. Метаданные в компьютерных системах // Программирование, МАИК/Наука «Интерпериодика». 2013. Т. 39, № 4. С. 28–46.

[2] Кравцов А.В. Информационные модели и технологии в организации работы научного сообщества по публикации и анализу коллекций исторических документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всероссийской научной конференции RCDL'2009. Петрозаводск: КарНЦ РАН, 2009. С. 210–218.

[3] Маркова Н.А. Логика биографических фактов //Информатика и ее применения, 2012. Т. 6, вып. 2. С. 49–58.

[4] Маркова Н.А. Программа Средства интеграции, хранения и анализа биографических данных (Фактограф). Свидетельство о государственной регистрации программы для ЭВМ № 2013617234 от 06.08.2013.

[5] Руководство пользователя по программе FamilySearch Indexing. © 2009, 2014 by Intellectual Reserve, Inc. URL: http://broadcast.lds.org/elearning/FHD/Local_Support/FamilySearchIndexing/RU/fsi_user_guide.pdf

[6] Столетие Московской 1-й гимназии. 1804-1904 гг. / сост. И. Гобза. – М.: Синод. тип., 1903. URL: http://dlib.rsl.ru/viewer/01003711731#?page=1

[7] Функциональные требования к библиографическим записям / Рос. библ. ассоц., РГБ. – М.: Пашков дом, 2008.

Formalization of the Fact-like Propositions in Specific Historical Studies

Natalia A. Markova

The paper proposes a model of metadata representation of the fact-like propositions that specify not only true statements, but suggestions, hypothesis, incomplete information, the results of analytic/synthetic processing. Requirements to provide efficiency of the specific historical studies are under consideration. The metadata are considered as the base of supporting IT.