Доступность конкретно-исторических данных в электронных коллекциях

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

Аннотация

Предложена мера доступности конкретно-исторических данных и рассмотрены факторы, на нее влияющие. Сформулированы требования к коллекциям, обеспечивающие эффективность информационно-логического доступа. Намечены перспективы усовершенствования коллекций.

1 Введение

Недалеко то время, когда содержание архивов и библиотек будет оцифровано и выложено на доступные для всех желающих электронные ресурсы. Целевые показатели единой системы информационно-библиотечного обслуживания ЛИБНЕТ [10] предполагают, что к 2015 году доля библиотечных фондов, переведенных в электронную форму, будет 50%, в том числе, библиотечных каталогов – 100%. Чуть медленнее, но все же достаточно интенсивно идет оцифровка архивов, по крайней мере, их справочно-информационного аппарата (СИА). Архивы приступают в плановом порядке к переводу в цифровой формат и представлению в Интернете наиболее востребованных архивных фондов по актуальной исторической тематике. В приоритетном порядке оцифровываются документальные комплексы, содержащие генеалогическую информацию, к которым существует устойчивый и широкий общественный интерес [6]. Таким образом, физическая доступность информационного богатства, накопленного в традиционных источниках, принципиально достижима. Сколько времени потратит пользователь на поиск ответа на свой вопрос в оцифрованных источниках, зависит от их информационно-логической доступности.

Особую социальную значимость приобрела в последнее время научно-познавательная деятельность, в которой изучается микроистория: история конкретной школы, конкретного села, конкретной семьи. Важнейшим фактором, способствующим повышению эффективности исследовательской деятельности, является возможность оперативного опубликования ее результатов в электронном информационном пространстве. Эти результаты становятся и новыми источниками и – при должном оформлении – новыми компонентами СИА.

Выделим два аспекта обеспечения логической доступности данных для задач микроистории. Первый – это наличие эффективной технологии выявления (поиска) источников по специфическим (неполным, неточным) запросам. Второй – наличие спектра средств представления, анализа и верификации конкретно-исторических данных. И в том и в другом случае эффективность реализации поддерживающего инструментария определяется уровнем систематизации конкретно-исторической информации, фиксацией метаданных (формализованного описания).

Как показала практика, основой систематизации целесообразно выбрать представление формализуемой части исторических данных, как компонентов объектно-ориентированной фактографической модели. Вариант такой модели – логика биографических фактов [7], [8] – был разработан и испытан на примере конкретного микроисторического исследования. Модель предполагает, что объекты исследования и сопряженные с ними объекты характеризуются временем существования, атрибутами и связями, которые также хронологически определены.

Определим, что влияет на доступность информации, с учетом специфики микроисторических исследований. Выбранная мера доступности – время, необходимое для получения нужных сведений. Перечислим факторы, влияющие на доступность традиционной (бумажной) и цифровой среды. На основе представления процесса доступа к информации, как последовательности взаимодействий пользователя и программно-технического инструментария, сформулируем требования к инструментальному оснащению электронных коллекций. Отметим необходимость фиксации содержательного (фактографического) представления, как базы для успешной реализации инструментария. Рассмотрим типовые примеры современных коллекций, содержащих конкретно-исторические сведения, оценим доступность содержащихся в них данных, а также сформулируем предложения по их развитию и совершенствованию.

2 Оценка доступности

2.1 Мера доступности

В качестве универсальной меры доступности выберем трудоемкость получения требуемой информации, которая, в свою очередь, выражается необходимым временем, и/или стоимостью информационных услуг, предоставляемых на возмездной основе. Для человека, самостоятельно выполняющего исследование, основная оценка – это время. Если часть работы требует денежных затрат, ее можно оценить во времени, необходимом для их получения.

Рассмотрим факторы, влияющие на доступность информации. Первая группа факторов – организационно-техническая – их мы просто перечислим, вторая – информационно-логическая будет рассмотрена подробнее.

2.2 Организационно-техническая доступность

В отечественной практике слово «доступ» ассоциируется со словом «секрет». «Секретность» – в общем случае, закрытость информации – требует времени и усилий на получение допуска, что нередко удлиняет время доступа до бесконечности.

Технические условия конкретного хранилища – «закрыто на ремонт», не имеет возможностей приема исследователей – могут отодвинуть момент получения документа на месяцы и годы. Регламент работы хранилища, связанный с его ресурсным обеспечением, диктует соответствующие задержки. Так, время выполнения заказа в Центральном историческом архиве Москвы – 5 рабочих дней. Если место доступа географически удалено, например, исследователь проживает в другом городе, придется добавить время (и материальные затраты) на поездку.

Документ в единственном экземпляре или в ограниченном их числе может быть выдан другому пользователю, и тогда, в случае традиционных архивов, время ожидания выдачи продлится еще на месяц.

Наконец, доступность документа может быть ограничена его состоянием: документ ветхий, требующий реставрации или особо ценный и на руки не выдается. Дополнительное время потребуется на его реставрацию/ изготовление копии. Если не предусмотрена платная услуга, задержка в этом случае может составить годы.

Перевод в электронную форму существенно облегчает положение, не устраняя, однако, до конца ни одной из перечисленных проблем. Даже географическая удаленность места доступа присутствует в тех библиотеках/ архивах, где доступ к электронным документам разрешается лишь из читального зала.

2.3 Информационно-логическая доступность в традиционных коллекциях

Если проблемы организационно-технической доступности решены, исследователю остается «только» найти нужные фрагменты нужных документов и извлечь из них необходимые сведения. Для решения задач поиска информации в традиционных информационных системах существовал вспомогательный набор средств – СИА.

СИА книги – это «совокупность дополнительных элементов издания, призванных пояснить, растолковать основной текст, способствовать усвоению содержания вошедших в издание произведений, облегчить читателю пользование изданием» [4]. Компонентами этого аппарата являются: оглавление, именные, географические, хронологические указатели, ссылки на используемые источники, а также, возможно, аннотации издания в целом и отдельных его частей.

Аналогичные, но более развитые составляющие характерны и для СИА хранилищ, которые включают всевозможные библиографические пособия, каталоги, картотеки, описи.

В рамках информационного универсума СИА – это совокупность библиографических и биобиблиографических изданий, путеводителей по архивам, справочных пособий.

Когда целевой документ найден, его нужно понять. Если он написан на иностранном языке, использует незнакомый пользователю набор терминов и сокращений, требуется изучить соответствующую кодовую систему или привлечь специалиста. То же относится, например, к рукописным текстам древнерусского письма.

Дополнительных усилий при чтении требуют плохая сохранность документа, витиеватый или небрежный подчерк, плохое форматирование текста, в частности, чересчур длинные абзацы. Грамотное использование средств выделения, шрифтов, интервалов, отступов и т.п. позволяет повысить доступность документа.

Перечисленные факторы влияния на доступность характерны и для понимания вспомогательных документов (индексов, каталогов и т.п. – компонентов СИА), которыми исследователь пользуется в процессе поиска целевого.

2.4 Информационно-логическая доступность в электронных коллекциях

С точки зрения микроисторических исследований, все документы, публикуемые в электронном виде, представляют собой либо в той или иной степени переработанные цифровые копии бумажных документов, либо публикации новых исследований, в конечном счете, опирающихся на бумажные документы (возможно, через цепочку,

включающую электронные варианты этих документов и другие исследования).

Реализация технологических преимуществ цифровой среды позволяет существенного повысить доступность конкретно-исторической информации. Для пользователя электронной коллекции переход к новому документу не связан с заполнением бумажного бланка-заказа и ожиданием его «ручного» выполнения, а переход к разделу книги с «ручным» листанием страниц.

Представим доступ к информации в цифровой среде, как процесс взаимодействия пользователя и программно-технического инструментария. Для этого введем следующие определения.

Назовем документальным объектом – документ, совокупность документов или фрагмент документа. Виртуальный документальный объект – представление на экране компьютера документального объекта и/или информации, сформированной автоматически из хранимых данных (например, базы данных). Часть виртуального документального объекта, видимую на экране в конкретный момент, назовем документальным окном. Документальное окно оснащается совокупностью элементов управления: гиперссылок, компонентов меню, линеек прокрутки, полей текстового ввода и пр., с помощью которых пользователь изменяет содержание документального окна.

Доступ к целевому документальному объекту складывается из последовательности шагов, на каждом из которых сменяется содержание окна, ознакомившись с которым, пользователь выбирает новое (уточняет) направление дальнейшего движения. Упрощенно: от пользователя требуется понять (интерпретировать) содержание текущего документального окна и сформировать задание на формирование следующего, от инструментального средства требуется его выполнить. Кроме того может потребоваться сохранение фрагмента текущего окна или его описания, а также действия по его сопоставлению с ранее полученной информацией.

Задания очередного шага связаны с продвижением окна по документальному объекту, поиском информации и/или ее представлением. Этот процесс в рамках традиционной технологии выглядел так: мы пролистывали книги; просматривали каталоги, индексы; обращались с листком-требованием к библиотекарю; использовали лупу или раскладывали на столе несколько страниц для их сличения.

3 Требования к инструментарию

Рассмотрим требования, каким должен обладать инструментарий электронной коллекции, чтобы обеспечить доступность конкретно-исторических данных, т.е. сократить трудоемкость, минимизировать время получения полезной информации.

Инструментарий должен обеспечивать различные варианты движения по документальному объекту. Топологические перемещения – окна по странице, между страницами, по крайней мере, не должны быть хуже, чем в бумажном первоисточнике.

Желателен прямой доступ по элементам оглавления и индексов, навигация по связям и отсылкам – гиперссылки.

Очевидным преимуществом электронных документов над их бумажными источниками, в том случае, если их текст распознан, является возможность полнотекстового поиска. В зависимости от обстоятельств, эта возможность реализуется от уровня конкретного документа или хранилища – до глобального поиска универсальными машинами. Однако для большинства областей, и, в том числе, конкретно-исторических исследований, этого недостаточно.

Объекты исследования, чаще всего, ищутся по именам. Имя понимается в широком смысле – это и название организации, и заглавие книги, и совокупность компонентов имени лица (в отечественной практике – личное имя, отчество, фамилия). Проблема в том, что между объектом и именем нет взаимно-однозначного соответствия. Рассмотрим несколько практических коллизий, в которых полнотекстовый поиск не дает удовлетворительных результатов.

Наличие тезок, в том числе, полных тезок, характерно не только для людей, но и для географических названий, наименований предприятий и т.п. Один и тот же объект часто известен под несколькими именами. Причем, имена не только изменялись во времени, но и нередко один объект одновременно назывался разными именами. Первая московская гимназия сначала называлась «губернской», затем «второй» («первой гимназией» в это время было совсем иное учебное заведение). Но и после переименования в «первую» в ходу были несколько имен, в частности, иногда ее называли «Первая мужская», потому как существовала и «Первая женская».

Вариативность – допустимость разных вариантов имен – вплоть до XX века была общей практикой. В настоящее время она, как правило, является результатом искажений. В коллекции «Подвиг народа» [2], содержащей сведения о награждениях защитников Отечества в Великой отечественной войне, находим:

Штурман Хайм Годелевич (1922, Винницкая обл., Козятинский р-н, с. Гуровцы)

Штурман Арсентий Григорьевич (1924, Винницкая обл., Казатинский р-н, с. Гурович)

Искажены не только отчества (у других братьев есть варианты «Гелеевич» и «Гордеевич»), но и названия села и районного центра.

Если в данной коллекции «Штурман» определено, как фамилия, то попытка использования этого слова в универсальной

поисковой машине приведет к колоссальному списку документов, относящихся к морякам, летчикам и автогонщикам.

Таким образом, в поисковом запросе микроисторического исследования хотелось бы указывать не абстрактное слово, но имя объекта определенного типа (его компоненты или варианты). Учитывая многозначность имен, весьма полезны уточнения области поиска, в частности, в виде хронологических и/или географических рамок

Важнейшими параметрами поиска кроме того являются атрибуты и связи объектов, также имеющие хронологические рамки. Чем более развитые средства формирования и исполнения поисковых запросов такого рода предоставляет инструментарий, тем эффективнее работа исследователя.

Промежуточные результаты поиска или очередного шага навигации – виртуальный документ, содержащий список отобранных документальных объектов. Эффективность работы с ним зависит от форм его представления. Сортировка, группировка (кластеризация), фильтрация результатов, а также различные виды инфографики (представление на схеме, географической карте, временной оси) – меры, способствующие повышению эффективности.

Найденный целевой объект также важно уметь рассматривать под различными углами зрения. Существует множество причин, требующих сохранения и совместного использование образа документа и его полнотекстового представления. Образ, при большом разрешении сканирования и применении цифровых «луп», позволяет разобрать тексты, ранее нечитаемые. Текстовое представление облегчает поиск и выборку цитат. Получаемое путем интерпретации фактографическое (содержательное) представление является основой верификации и анализа данных. Вопрос об его фиксации, в том числе, для повторного использования другими пользователями, представляет значительный интерес.

4 Фактографическое представление

Для того чтобы обеспечить перечисленные требования к инструментарию, нужно не только распознавание текста документов, но и/или распознавание их смысла – фактографическое индексирование. То, каким образом получаются, хранятся, используются его результаты, существенным образом зависит от применяемой информационной технологии. Смысловое или фактографическое представление документального объекта – метаданные – в зависимости от используемой технологии включается, как компоненты разметки в текст документа, и/или хранится отдельно в виде индексов, каталогов, справок.

Для конечного пользователя модель метаданных не видна, он пользуется функциональностью инструментария. За счет более сложных алгоритмов можно в какой-то мере покрыть недостатки представления, или, что случается чаще, не использовать потенциальные возможности, заложенные в модели данных. В этом случае налицо перспектива усовершенствования. В целом, если оставить в стороне наиболее болезненные вопросы о полноте метаданных и их соответствии документальным объектам, то эффективность доступа в значительной степени определяется используемой концептуальной моделью метаданных.

Формализуемая часть задач микроистрии может быть сформулирована в терминах фактов, касающихся изучаемых объектов, их свойств и взаимосвязей. Целесообразно основывать модель представления фактов на объектно-ориентированном подходе. Сведения об объектах рационально привязывать к временной оси: каждому объекту сопоставлять хронологические рамки его существования, его атрибуты и связи также соотносить со временем. Это требование существенно отличается от положений библиографических и генеалогических стандартов, где временные характеристики определяются, как отдельные атрибуты объектов. При этом на практике ему удовлетворяют некоторые электронные коллекции.

В работах [7], [8] объектно-ориентированная темпоральная модель была определена формально и подробно сопоставлена с имеющимися библиографическими и генеалогическими стандартами. Единицы модели названы фактами. Модель дополнительно предусматривает возможность формулировки утверждений, сопоставляющих факты логически и темпорально. Например, «верен один из альтернативных фактов»; «если .., то…»; «событие … предшествует событию …».

Номенклатура классов объектов, их отношений и связей – важнейшая характеристика модели. Избавиться от неоднозначности имен, представить связную картину микроистории возможно, если вместо текстовых атрибутов будут выделены объекты и указаны их связи.

Рассмотрение документальных объектов, как специального класса объектов, наравне с классами «организация» или «персона», позволяет единообразно относиться к фактографической и библиографической составляющим метаданных.

К сожалению, рассчитывать на автоматическое распознавание смысла для подавляющего большинства исторических документов в обозримом будущем не приходится. Трудовые ресурсы профессионалов (историков, архивистов, библиографов) весьма ограничены. Для того чтобы объединить усилия профессионалов и задействовать колоссальные ресурсы непрофессионалов и при этом чтобы результат их работы был систематизирован и допускал верификацию, требуется специальная технология. Позитивные

примеры технологий, обеспечивающих привлечение добровольцев для фактографического индексирования, будут приведены в следующем разделе.

5 Доступность современных коллекций и потенциал ее повышения

Рассмотрим, как производится доступ к данным в современных электронных коллекциях конкретно-исторической направленности с точки зрения перечисленных требований. Начнем рассмотрение со «слабых» в точки зрения фактографического наполнения и функциональности инструментария примеров. Затем представим передовые технологии, потенциал развития которых также далеко не исчерпан.

5.1. Собрание славянских рукописей из архивов Российской государственной библиотеки и Московской духовной академии [12] – это набор из около 5000 рукописей, представленных файлами-образами страниц. Для чтения письма XVI-XVIII веков требуются знания из области палеографии. СИА – далеко неполные кратчайшие описания отдельных документов. Очень ценная и очень малодоступная с логической точки зрения коллекция. Перспектив улучшений множество – начиная с возможности переходов между последовательными страницами одного документа.

5.2. Российский государственный архив древних актов провел оцифровку более сотни описей и выложил их на своем сайте [11]. Описи – это образы машинописных (плохого качества) страниц. Искать что-либо в такой описи можно только путем их сплошного просмотра (с небольшой помощью оглавления). Найденная информация – архивный шифр, с которым нужно обращаться в архив за документом. Требуется, как минимум, распознавание текстов – описей – перевод их в полнотекстовый вид.

5.3. В ресурсе общества «Мемориал» [5] представлены краткие биографические справки репрессированных лиц. Обеспечены навигация по алфавиту имен и полнотекстовый поиск. Индексирования по времени/месту рождения/ смерти, виду деятельности не предусмотрено. Не выделены, как объекты, места заключения и документы-источники. Тексты справок представляют собой редкий случай исторических источников с четко определенной структурой. Это именно тот случай, когда имеется возможность автоматизированного фактографического индексирования. При этом рационально воспользоваться объектно-ориентированной темпоральной моделью представления фактов.

5.4. Международный проект – Всемирное семейное древо [1] – один из крупнейших и наиболее четко организованных в мире. На 13.05.2013 число индексированных записей в нем составляло 1,010,740,457. Коллекция оснащена аппаратом, включающим структурно-упорядочивающие формы ввода, поиска и просмотра информации. Редчайший случай среди современных ресурсов – поиск осуществляется по диапазону дат выбранной категории события. В коллекции очень строго соблюдается принцип опоры на источники. Фактографическая информация всегда содержит отсылку на первоисточник. Колоссальный объем ресурса – результат деятельности более 150 тысяч волонтеров, выполняющих фактографическое индексирование. Однако спектр объектов изучения в ресурсе только люди, и только родственные отношения между ними фиксируются. Ограничена также номенклатура первоисточников. Таким образом, для задач микроистории эта коллекция имеет ограниченное применение.

5.5. Коллекция Родовод [9] во многом схожа с FamilySearch – здесь, также центром рассмотрения являются люди и их родственные связи. В Родоводе нет строгой дисциплины отслеживания источников, что, с одной стороны, множит возможные искажения, с другой, позволяет оперативно набросать эскиз исторической картины, опираясь на произвольные, в том числе, устные источники. Используемый в Родоводе принцип wiki, предполагающий коллективный контроль и редактирование, позволяет в какой-то мере бороться с дефектами. Ресурс информационно открыт, принципиально наращиваема номенклатура событий. В информационное представление объектов Родовода включаются форматированные тексты, оснащаемые гиперссылками и другими компонентами wiki-разметки, что позволяет сочетать формализованное и неформальное знание.

К сожалению, в Родоводе недоработан механизм поиска, который на базе уже имеющегося представления вполне может обеспечить поиск по временным диапазонам и связям. Несложно, в виде дополнительных модулей, реализовать средства контроля вводимых данных (например, сопоставить время жизни родителей – детей). За счет доработки, в Родоводе могут быть добавлены объекты, отличные от генеалогических.

5.6. Википедия [3] рассматривает объекты всех категорий, не только исторические. В некотором смысле ее статья – это вывернутый наизнанку объект Родовода. Здесь формализованные, содержащие метаданные, фрагменты вкраплены в неформальный текст. Потенциал повышения доступности Википедии далеко не исчерпан. Так, вместо простой гиперссылки целесообразно использовать ее семантическое расширение, раскрывающее характер связи (например, посредством пометки «друг», «отец»), что уже предполагают современные микроформаты. Другое необходимое расширение, которое, к сожалению, пока даже не декларируют создатели семантического web (но вполне представимо, как его развитие), – это определение динамики связей. Привязка событий и связей к временной оси – важнейшее условие успешности конкретно-

исторического поиска. Тогда, то, что некто жил в таком-то городе в такое-то время, позволит не включать его в список потенциальных соседей целевого персонажа, жившего там в другое время. Внедрение пометок-спецификаций связей способно существенно улучшить категоризацию статей. Связь «учился в» Первой гимназии сразу определит определенную классификационную категорию, что существенно облегчит навигацию по коллекции. Для этого, правда, придется ввести параметризованные категории – в приведенном примере от учебного заведения – Первой гимназии.

5.7. В инструментальном комплексе Фактограф [8] была реализована предложенная в [7] модель представления биографических и биобиблиографических данных. Модель предлагает не только явную хронологически определенную фиксацию связей между людьми и объектами произвольной природы, но и возможность работы с многозначными терминами, искаженными и дефектными данными. При этом в рамках конкретного применения комплекса объявленная «биографическая» направленность может быть изменена на произвольное направление – краеведение, историю организации и пр. Залогом этому – открытость номенклатуры классов объектов, их атрибутов и связей.

Прототипная реализация предложенной концепции и использование Фактографа в конкретно-историческом исследовании продемонстрировали его высокую эффективность для решения задач интеграции и аналитико-синтетической обработки фактов, получаемых из гетерогенных источников. В настоящее время Фактограф предназначен для индивидуального исследователя. В перспективе стоит задача организации коллективной работы.

6 Заключение

Для того чтобы информация из конгломерата оцифрованных документов была бы действительно доступна, требуются значительные усилия, в том числе, связанные с привлечением трудовых ресурсов. Существует армия любителей истории, которые могли бы приложить свои силы к фактографическому индексированию источников. На тысячах площадках – в социальных сетях и специализированных сайтах – миллионы пользователей обмениваются данными и ссылками, предоставляют документальные свидетельства общему доступу. Информация, как правило, столь плохо систематизирована, что ее использование чрезвычайно затруднено.

Усовершенствование инструментария передовых коллекций вместе с созданием новых технологий, основанных на объектно-ориентированном подходе, позволит (с привлечением усилий виртуальных сообществ) существенно повысить доступность конкретно-исторических данных. Возможные пути развития и концептуальные основы таких инструментов были представлены в настоящей работе.

Литература

[1] FamilySearch /Intellectual Reserve, Inc. 2013. https://familysearch.org/

[2] Банк документов «Подвиг народа в Великой Отечественной войне 1941-1945 гг.». / Сайт Министерства обороны Российской Федерации. http://www.podvignaroda.mil.ru/

[3] Википедия – свободная энциклопедия /Wikipedia® http://www.wikipedia.org

[4] ГОСТ Р 7.0.1-2003 «СИБИД. Издания. Знак охраны авторского права. Общие требования и правила оформления».

[5] Жертвы политического террора в СССР /Общество «Мемориал». http://lists.memo.ru/

[6] Концепция развития архивного дела в Российской Федерации. http://archives.ru/documents/projects-concept-razvitie-archivnogo-dela.shtml.

[7] *Маркова Н.А.*. Логика биографических фактов // «Информатика и ее применения», 2012. Т. 6. Вып. 2. С. 49-58.

[8] *Маркова Н.А*. Электронная коллекция биографических фактов. //Труды 14[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2012, Переславль-Залесский, Россия, 2012. С. 287-293.

[9] Многоязычное генеалогическое древо. http://www.rodovid.org/

[10] Основные направления развития Общероссийской информационно-библиотечной компьютерной сети ЛИБНЕТ на 2011-2020 годы. – М.: 2011. http://www.nilc.ru/nilc/documents/libnet-2011-2020.pdf.

[11] Сайт Российского государственного архива древних актов. http://rgada.info/

[12] Славянские рукописи. //Сайт Свято-Троицкой Сергиевой лавры. http://old.stsl.ru/manuscripts.

Accessibility of Concrete Historical Data in Digital Collections

Natalia A. Markova

The paper defines a meager to estimate accessibility of concrete historical data. It considers impacts of different factors on accessibility. Tool requirements to provide efficiency of information logical access to digital collections are under consideration. The paper suggests prospects of improvements for several typical modern collections.