Использование связанных данных для улучшения тематического доступа в онлайновых первоисточниках - тематическое исследование он-лайновой коллекции о I Мировой войне в Университете Колорадо, Баулдер

Теа Линдквист (Thea Lindquist)
Университет Колорадо, Баулдер
(University of Colorado Boulder)
Boulder, CO, USA
E-mail: thea.lindquist[at]colorado.edu

Эро Хивонен (Eero Hyvönen)

Юха Торнроос (Juha Törnroos)

Эту Макела (Eetu Mäkelä)
Университет Аалто и Университет Хельсинки
(Aalto University and University of Helsinki)
Aalto, Finland
E-mail: firs.last[at]aalto.fi

Перевод на русский язык:
Елена Загорская
(Российская национальная библиотека, Санкт-Петербург)

Session: *117* — **Subject access now: inspiring, surprising, empowering** — *Classification and Indexing*

Аннотация:

Академические пользователи часто считают работу с доступными online первичными источниками и сложной, и благодарной. Совершенствование тематического доступа к этим источникам имеет важное значение для пропаганды цифровых коллекций, работа с первоисточниками становится все более важной в учебных планах гуманитарных наук. В Университете Колорадо (Боулдер) было проведено исследование потребностей пользователей-гуманитарев и их опыта работы с этими источниками. Были изучены две основные потребности, результатом исследования стало повышение находимости источников и раскрытие их контеста, особенно для источников по исторической тематике.

Связанные данные способствуют удовлетворению этих потребностей, связывая понятия из источникав с терминами специализированного словаря, обогащая их внешними ресурсами, что обеспечивает большее количество семантических услуг для

пользователей. В данном докладе предлагается обсудить авторское исследование, предприняое для улучшения тематического доступа к онлайновой электронной коллекции Университета Колорадо по Первой мировой войне с глубокими связями исторических данных о жизни гражданского населения в оккупированной Бельгии. Интернет глубокие связи исторических данных о гражданском опыт в оккупированной Бельгии. Эта работа призвана обогатить наше понимание движущих сил в период Первой мировой войны.

I. ВВЕДЕНИЕ

С ростом количества онлайновых коллекций цифрового культурного наследия, пользователи имеют больше возможностей, чем когда-либо прежде, для прямого доступа к первоисточникам.[1] Однако, академические пользователи считают работу с этими источниками не только полезной, но и сложной задачей. Хотя эти электронные коллекции предлагают весьма ценные ресурсы, особенно по гуманитарным дисциплинам, исследования показывают, что они по-прежнему недостаточны. Следовательно, улучшение доступа для этой целевой группы академических пользователей, приобретает все более важное значение.[2] Для того чтобы понять их потребности и облегчить их работу с источниками, было проведено исследование потребностей пользователей и студентов гуманитарного факультета Университета Колорадо, Боулдер (University of Colorado Boulder (CU)). К двум основным выявленным требованиям отнесены: улучшение находимости и контекст, в частности, на исторические темы.

В докладе рассматривается, как связанные данные[3] могут удовлетворить эти требования при использовании онлайн-коллекции этого университета о Первой мировой войне, ставшей экспериментальной базой исследования. В дополнение к представлению метаданных коллекции в качестве связанных данных, нашей целью является изучение глубокой связи исторических данных в источниках, связанных с гражданской жизнью в оккупированной Бельгии, чтобы показать все сложные вопросы, на которые можно ответить с помощью автоматизированных методов, применяемых в специализированном домене. Этот подход помогает удовлетворить потребности пользователей, связывая понятия из источников со специализированными словарями, обогащая их с помощью внешних ресурсов, а также позволяет предоставить пользователям семантически более богатые сервисы.

II. ПРОБЛЕМЫ работы с первоисточниками онлайн

Работа с первичными источниками становится все более важным компонентом в гуманитарных науках, и особенно в истории, в учебных программах для студентов и в средних учебных заведениях.[4] Действительно, использование онлайн-первоисточников в классе считается основополагающим принципом в современных педагогических методиках, поскольку они способствуют развитию критического мышления и конструктивной методики обучения. Lee и Clarke, например, объясняют, что «нелинейная форма Web может стать инструментом поощрения студентов к работе с несколькими последовательностями, голосами, результатами и последствиями исторического повествования"[5]. Онлайновые первоисточники представляют различные преимущества для исследований по сравнению с нецифровыми форматами, главным образом в том, что они более доступны, возможен их поиск, их форматы более гибкие, ими легко манипулировать.

Несмотря на то, что коллекции оцифрованы и доступны в учреждениях культуры, и это наследие представляет чрезвычайно богатые и ценные материалы для преподавания, обучения и исследований в гуманитарных науках, они по-прежнему недоступны онлайн, как для преподавателей, так и для студентов по целому ряду причин. Поскольку

большинство из этих пользователей полагаются на Google, деконтекстуализация источников, непрозрачность баз данных учреждений являются, в результате, препятствиями при их использовании. Учитывая проблемы с находимостью и доступностью цифровых первоисточников, вряд ли стоит удивляться, что некоторые исследования выявили повышение спроса на более простой тематический поиск документов и коллекций в этих базах данных.[6]

III. Оценки потребностей пользователей

При разработке ориентированного на пользователя цифровых образовательных ресурсов инструмента для работы с онлайн-первоисточниками, в Университете Колорадо в январе 2011 года было проведено исследование потребностей пользователей.[7] Целью данного исследования было понять потребности академических (университетских) пользователей в области гуманитарных наук и облегчить их взаимодействие с этими источниками, основываясь на прямой обратной связи с ними. Базой исследования стали более 20 полуструктурированных интервью с преподавателями, аспирантами и студентами, представляющими широкий спектр образовательных уровней и дисциплинарных областей, а также различную степень знакомства с первоисточниками.[8] Несмотря на то, что основное внимание уделялось потребностям академических (университетских) пользоватей, академические потребности, которые они выразили, характерны, в той или иной мере, для всех пользователей онлайновыми первоисточниками.

Исследование подтвердило, что преподаватели и студенты гуманитарных факультетов по-прежнему сталкиваются с серьезными проблемами при поиске и контекстуализации онлайн-первоисточников.[9] Они, как правило, не воспринимают весь спектр имеющихся ресурсов и считают неэффективным поиск в нескольких базах данных и веб-сайтах для выявления релевантных источников. Как только они определяют коллекцию для поиска, они сталкиваются с проблемами нахождения и контекстуализации отдельных источников и содержащихся в них сведений, особенно по исторической тематике.

Участники анкетирования сообщили, что библиографические метаданные часто недостаточны для предоставления доступа к отдельным источникам и, особенно, к их разделам по темам, хронологии и географическим районам с нужной степенью детализации. Так как похожие понятия отражаются в текстах по-разному (разными терминами), поиск по ключевым словам текста является случайным. Онлайновые первоисточники еще более восприимчивы к деконтекстуализации, поскольку поисковое ключевое слово стимулирует пользователей искать фрагмент документа, в котором данный термин упоминается, и переходить к следующему фрагменту, а не к чтению документа в полном объеме.[10] Кроме того, поисковые машины и системы ссылок к онлайн-источникам могут добавить к данной проблеме отделение отдельных документов от архивов, из которых они происходят.

Контекстуализация первичных источников часто бывает необходима, чтобы сделать их достаточно доступными для пользователей, особенно для студентов и неспециалистов, взаимодействовать с субстанцией материала. Контекст может включать в себя отображение взаимосвязей между отдельными документами, а также ресурсы, которые помогают объяснить, как каждый документ и информация из него вписываются в соответствующий исторический контекст.[11] Даже используя релевантные источники и адекватный контекст, пользователи могут сталкиваться с дополнительными проблемами исследования первосточника: иностранные языки, предвзятость документа, исторические обычаи, орфография, грамматика, палеография/типография и т.д.[12] Хотя все эти проблемы затрудняют понимание и использование онлайн-первоисточников, участники исследования согласились с тем, что эти источники представляют уникальные образовательные и исследовательские возможности.

IV. Преимущества связанных данных

При оценке различных методов, которые способствовали бы удовлетворению потребностей этих пользователей, в целях повышения совместимости и эффективности использования цифровых исторических коллекций одним из самых перспективных методов было признано использование связанных данных и их семантически богатых сервисов. По словам Тома Хита (Tom Heath) и Кристиана Бизера (Christian Bizer), связанные данные "относятся к ряду наилучших способов публикации и сопряжения структурированных данных в Web."[13] При установлении связей между родственными понятиями внутри и между документами таким образом, что они становятся понятными для компьютеров, связанные данные позволяют (1) агрегатировать данные из распределенных онлайн первоисточников, (2) создавать новые связи между ними и визуализировать их таким образом, который ранее был недоступен, и (3) обогащать данные с помощью ссылок на внешние ресурсы, такие как DBpedia.[14]

Содержание цифрового культурного наследия и исторические материалы, в особенности, представлют такой уровень сложности, для которого очевидны преимущества семантического обогащения (мета)данных и интеллектуальных пользовательских сервисов, повышающие находимость и обогащающие контекст.[15] Они выявляют сложные, часто нелинейные связи между темами, людьми и местами, скрытые в источниках, особенно, если эти методы основываются на онтологиях и других специализированных словарях, которые придают смысл понятиям и отношениям между ними в конкретном историческом аспекте. Приложения связанных данных также могут способствовать решению распространенных проблем формулирования запросов с использованием предметных рубрик и поиска по ключевым словам в исторических источниках, например, предлагая пользователям поисковые термины, которые повышают эффективность и точность поиска. Таким образом, они способствуют повышению эффективности тематического доступа к историческому контенту.

Кроме того, связанные данные способствуют более богатой контекстуализации источников при подключении связей, представленных не только в коллекциях, но и с релевантными внешними источниками, что также повышает взаимодействие, обмен и повторное использование данных из исторической коллекции. Кроме того, связанные данные могут представлять организационную структуру коллекций таким способом, который не только сохраняет первоначальный контекст документов, но и предупреждает пользователей о доступных типах материалов. Взятые вместе, эти усовершенствования помогут преодолеть известные ограничения поиска по предметным рубрикам и неточность поиска по ключевым словам, позволяя сравнивать исторические сведения во времени и пространстве, а также подключая к работе сразу несколько коллекций. Кроме того, способ представления связанных данных легко осваивается и способствует развитию интеллектуальных приложений, которые просто перестраиваются и предоставляют пользователю широкий выбор возможностей для анализа и визуализации данных.

V. ПРОЕКТ СВЯЗАННЫХ ДАННЫХ О ПЕРВОЙ МИРОВОЙ ВОЙНЕ

К этому проекту привлечены ученые, специализирующиеся в области компьютеризации из Университета Аалто[16] и библиотечные специалисты-эксперты в области тематического индексирования из Университета Колорадо. В качетстве первичного массива данных использовалась электронная коллекция документов Первой мировой войны, представленная на сайте Университета Колорадо, которая содержит более 1100 названий (55.000 страниц), опубликованных с 1829 по 1922 год, преимущественно между 1914 и 1918 гг.[17] Происхождение коллекции не совсем ясно, вероятно, она вошла в фонды библиотеки Университета Колорадо в 1920-1930-х годах в связи с работой над проектом «Колорадо в Первой мировой войне», которую проводил профессор истории Джеймс

Филд Уиллард (James Field Willard), собиравший документы о деятельности граждан и государственной деятельности во время войны.[18] Документы коллекции происходят, в основном, из США и касаются различных геополитических регионов и проблем, от этнических и религиозных конфликтов до проблем империи и колоний. В коллекции представлен широкий диапазон жанров, в том числе брошюры, книги, доклады, выступления и карты. В настоящее время ведутся переговоры о опубликации этих материалов в виде связанных открытых данных в соответствии с лицензией некоммерческой организации Creative Commons 2.0.

Одной из основных целей проекта является улучшение тематического доступа к онлайн-коллекции и формирование контекста документов путем установления связей между точками данных в коллекции, базами данных, включенных в проект, и внешними источниками данных, таких как DBpedia и Freebase.[19] Другая цель заключается в содействии аннотированию и установлению глубоких связей между понятиями из коллекции Первой мировой войны в специализированном историческом поддомене, в данном случае, относящихся к жизни гражданского населения оккупированной Бельгии во время Первой мировой войны. Эта тема была выбрана не только потому, что она хорошо представлена в коллекции, но и потому, что влияние «тотальной войны» на гражданское население является предметом современных научных исследований. Большинство публикаций, относящихся к этой категории, касаются трудностей бельгийцев, пострадавших во время немецкого вторжения и оккупации, в частности, зверских инцидентов, таких как убийства и депортация работников и последствия военного правления в ежедневной жизни. Среди данных, преобразованных к настоящему времени в RDF (Resource Description Framework - Формат Связанных данных), представлены: коллекция метаданных (MARC), стандартные словари Имперского военного музея (Imperial War Museum: IWM)[20], сведения о немецких зверствах в Бельгии, и немецкая армейская иерархия.[21]

Глубокая система связей, использующая специализированные словари по Первой мировой войне в Бельгии в сочетании с интеллектуальным пользовательским интерфейсом, разработаны для демонстрации таких типов сложных вопросов, на которые можно ответить в этом поддомене при удовлетворении таких потребностей пользователей, как: Отражены ли в коллекции литературы масштабы зверств, совершенных Немецкими войсками в Бельгии? Какие подразделения Немецкой армии были вовлечены в большинство инцидентов? Каково было географическое распределение депортаций из бельгийских провинций? Такой тип функциональности должен приводить к большему пониманию многих сил, формирующих период Первой мировой войны. Учитывая весьма специфическую природа этого субдомена и отсутствие готовых онтологий, эти словари должны были быть созданы путем адаптации терминологии и конструкций из стандартной печатной библиографии о Бельгии в Первой мировой войне[22], включая обратную связь от историков данной темы[23], и связывая везде, где возможно, релевантные термины из других наборов данных, включенных в проект, например, из коллекции метаданных и словарей IWM.

В нашей работе используется существующая система онтологий FinnONTO[24], расширенная аннотациями данных, упомянутых выше, и общей моделью основных событий Первой мировой войны с использованием семантических инструментов аннотирования SAHA.[25] Кроме того, для автоматизации части процесса аннотирования мы использовали инструмент ARPA[26]. ARPA является инструментом извлечения информации, которая автоматически распознает в текстовых документах имена лиц и ключевые понятия из онтологии. Предлагаемые аннотации могут быть проверены и исправлены вручную с помощью редактора SAHA. Наконец, для улучшения поиска и просмотра данных по темам, людям, местам и периодам времени, и для представления его

в визуальной и интерактивной формах ведется разработка специализированного web-портала Первой мировой войны на основе поисковика фасетизированного портала HAKO.[27] Модель портала Первой мировой войны и другие структуры, которые мы создали, предназначены для совместного использования, что обеспечивается тем самым «семантический клеем", который связывает отдельные базы данных вместе и позволяет производить поиск и просмотр между ними. Кроме того, стратегия, которую мы разработали для этого проекта, может быть адаптирована к другим историческим доменам и базам данных, в частности, связанных с такими конфликтами, как Гражданская война в США, или Вторая мировая война.[28]

VI. ЗАКЛЮЧЕНИЕ

При установлении связей между понятиями в базах данных по Первой мировой войне с использованием специализированных словарей, позволяющих представлять семантически обогащенные сервисы, мы надеемся предоставить пользователям возможность осуществлять квалифицированный онлайновый поиск и эффективно использовать первоисточники. Приближающаяся столетняя годовщина этой войны, несомненно, вызовет активизацию интереса к ее истории, особенно в странах, которые были ее участниками. Мы можем использовать этот момент, чтобы привлечь внимание пользователей к активному изучению прошлого и богатству цифровых материалов, которые учреждения культурного наследия сделали доступными.

Благодарности. Эта работа была поддержана Фулбрайтовской программой научных исследований США (Fulbright U.S. Scholar Program), Финским фондом культуры (Finnish Cultural Foundation), и Финским агентством по финансированию технологий и инноваций (Tekes - Finnish Funding Agency for Technology and Innovation).

---

[1] Первичными источниками являются документы, предметы или другие свидетельства прошлого, которые были созданы в течение того периода времени, когда имели место исторические события, или тех людей, которые пережили эти события. Например: дневники, письма, выступления, правительственные документы, книги, интервью, фотографии, аудио-и видеозаписи, и артефакты.

[2] D. Harley, “Use and Users of Digital Resources: A Survey Explored Scholars Attitudes about Educational Technology Environments in the Humanities”, *Educause Quarterly* 30, no. 4 (2007): 12-20.

[3] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the Semantic Web: Theory and Technology, ed. J. Hendler and F. van Harmelen (San Rafael, CA: Morgan & Claypool, 2011). Freely available at: http://linkeddatabook.com/editions/1.0/.

[4] Смотри, например:
J.K. Lee, “Digital History and the Emergence of Digital Historical Literacies”, in *Technology in Retrospect: Social Studies in the Information Age, 1984-2009,* ed. R. Diem and M.J. Berson (Charlotte, NC: Information Age Publishing, 2010), 78-80.
D. Malkmus, “‘Old Stuff’ for New Teaching Methods: Outreach to History Faculty Teaching with Primary Sources”, *portal: Libraries & the Academy* 10, no. 4 (2010): 414-416.

[5] J.K. Lee and W.G. Clarke, “High School Social Studies Students’ Uses of Online Historical Documents Related to the Cuban Missile Crisis”, *Journal of Interactive Online Learning* 2, no. 1 (2003): 3.

[6] M.C. Pattuelli, “Modeling a Domain Ontology for Cultural Heritage Resources: A User-Centered Approach”, *Journal of the American Society for Information Science & Technology* 62, no. 2 (2011): 314-342.

[7] Университет Колорадо классифицируется как Исследовательский университет Карнеги (очень высокий уровень исследовательской работы) с широким диапазоном грантов для магистерских и докторских исследований в гуманитарных науках.

[8] Участники были включенны в число семи гуманитарных колледжей и ведомств на территории кампуса: архитектура и планирование; исследования классических, английского, французского и итальянского языков; исторические исследования; исследования музыки; и религиоведение. Каждое из направлений предлагает программы на докторском уровне.

[9] Это резюме не включает в себя полный спектр потребностей пользователей, выявленных в ходе исследования, а только те, которые относятся к теме данной статьи. Для полного обсуждения смотри: T. Lindquist and H. Long, “How Can

Educational Technology Facilitate Student Engagement with Online Primary Sources?: A User Needs Assessment", *Library Hi Tech* 29, no. 2 (2011): 224-241.

[10] J. Garrett, "KWIC and Dirty? Human Cognition and the Claims of Full-Text Searching", *Journal of Electronic Publishing* 9, no. 1 (2006), available at: http://quod.lib.umich.edu/cgi/t/text/textidx?c=jep;cc=jep;q1=garrett;rgn=main;view=text;idno=3336451.0009.106 (accessed 13 February 2012).

[11] Одна студентка привела следующий пример типа контекста, который она хотела бы видеть: "Я хотела бы больше данных о фоне и контексте первоисточников, с которыми я работаю. [Многие онлайн коллекции первоисточников] только представляют источник, но не передают смысла того письма, например, которое было доставлено в разгар вспышки холеры" (Lindquist and Long, 233).

[12] T. Lindquist and H. Wicht, "Pleas'd By a Newe Inuention?: Assessing the Impact of Early English Books Online on Teaching and Research at the University of Colorado at Boulder", *The Journal of Academic Librarianship* 33, no. 3 (2007): 347-360.

[13] Heath and Bizer, chap. 2, "Principles of Linked Data", accessed 23 May 2012: http://linkeddatabook.com/editions/1.0/#htoc8.

[14] DBpedia является версией Википедии, представленной в виде связанных данных (http://www.dbpedia.org/).

[15] E. Hyvönen, "Semantic Portals for Cultural Heritage", in *Handbook on Ontologies*, 2d ed., ed. S. Staab and R. Studer, International Handbooks on Information Systems (Berlin: Springer, 2009).

[16] Semantic Computing Research Group, смотри http://www.seco.tkk.fi/.

[17] Смотри http://libcudl.colorado.edu/wwi/index.asp.

[18] Эти материалы стали основой Исторической коллекции Университета Колорадо и Архива Университета (David M. Hays, "The History of the Archives, University of Colorado at Boulder Libraries, 1917-2011" [неопубликованный доклад, Archives, University of Colorado Boulder Libraries], 1-2).

[19] Смотри http://www.freebase.com/.
Freebase — большая база знаний, содержащая метаданные, собранные, в основном, Интернет-сообществом из множества источников, например, из отдельных вики-проектов. Целью Freebase является создание глобального ресурса, который позволит людям (и машинам) иметь более эффективный доступ к общеизвестной информации. Разрабатывается американской софтверной компанией Metaweb и работает публично с марта 2007.

[20] К ним относятся принятые ключевые слова из названий событий, связанных с Первой мировой войной, утвержденные (принятые) географические ключевые слова, связанные с Западным фронтом, из таксономии Getty TGN, дополненные терминами, связанными с коллекцией IWM, и таксономией IWM по тематике Первой мировой войны. Выражаем благодарность Имперскому военному музею за предоставленную возможность совместного использования этих словарей.

[21] Отдельное спасибо Джону Хорну (John Horne) и Алану Крамеру (Alan Kramer) из Тринити-колледжа Дублина, которые собрали и проанализировали эти злодеяния и дали разрешение включить их материалы в проект (J. Horne and A. Kramer, *German Atrocities, 1914: A History of Denial* [New Haven: Yale University Press, 2001], Appendix 1, 435-439).

[22] P. Lefèvre and J. Lorette, eds., *La Belgique et la Première Guerre mondiale: Bibliographie*, 2 vols. (Brussels: Musée Royal de l'Armée, 1987-2001).

[23] Большое спасибо Марте Ханна (Martha Hanna, University of Colorado Boulder), Софии де Шёпдрийвер (Sophie de Schaepdrijver, Pennsylvania State University) и Темми Проктору (Tammy Proctor, Wittenberg University) за их предложения.

[24] Смотри http://www.seco.tkk.fi/projects/finnonto/.

[25] Коды загруженных публикаций и источников смотри на http://www.seco.tkk.fi/services/saha/.

[26] Смотри http://www.seco.tkk.fi/services/arpa/.

[27] Смотри http://www.seco.tkk.fi/tools/hako/.

[28] Характер содержания облегчает установление связей на основе военных концепций и структур, таких как рода войск, полков и сражения.