

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript-2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12-06-06061-г),
гранта РГНФ (проект № 12-04-14154-г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:

В. А. Баранов, д-р филол. наук, проф.

А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

Информационные технологии и письменное наследие [Текст] :
И741 материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября
2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ;
Ижевск, 2012. — 328 с.

ISBN 978-5-8021-1402-5

Сборник содержит материалы конференции, посвященной современ-
ным электронным средствам хранения, описания, обработки, исследования
и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978-5-8021-1402-5

© Петрозаводский государственный
университет, 2012
© Ижевский государственный технический
университет им. М. Т. Калашникова, 2012

РЕПРЕЗЕНТАЦИЯ СОДЕРЖАНИЯ ИСТОРИЧЕСКИХ ИСТОЧНИКОВ В СЕМАНТИЧЕСКИХ СЕТЯХ: АТТЕМПТО CONTROLLED ENGLISH (ACE)

А. С. Иванов¹, А. Г. Варфоломеев²

¹Институт регионалистики (REGI), Резекненская высшая школа, Резекне, Латвия;

²Петрозаводский государственный университет, Петрозаводск

The paper discusses some problems of representation of the tenor of historical records in semantic publications. The authors argue that the tenor of historical records can be fully represented on the Semantic Web by means of Attempto Controlled English (ACE). The advantages of this natural controlled language are rather obvious: ACE texts can be translated into the so-called Discourse Representation Structures (DRS) that can be directly correlated with first-order logic formulas. Furthermore, documents' texts in ACE can be processed using reasoner, which generates new hypotheses on the bases of the facts revealed by a researcher. ACE makes it also possible to record axioms and inference rules, e.g. "if something X is written by someone Y, then Y is the author of X". As a result, a great number of facts related to the texts, which should be entered and processed manually in the course of production of the semantic publications on the basis of Semantic Me-diaWiki, can be entered and processed automatically due to the opportunities offered by ACE. At the same time, new knowledge about the texts of the written historical records, namely new facts and hypotheses, can be acquired by means of automatic inference. Therefore, ACE seems to be a very promising tool for the purposes of the semantic publication of historical records.

В научной литературе и в практике подготовки Web-изданий (преимущественно — электронных научных журналов) и цифровых библиотек [Baruzzo et al., 2009; Shotton, 2009] под семантической публикацией обычно понимается электронная публикация текста, снабженная дополнительным информационным слоем, который представляет знание о тексте, а также смысл текста в формализованном виде [Ahonen and Hyvönen, 2009; Mirzaee et al., 2005]. Общеизвестно, что семантические публикации не только существенно расширяют возможности поиска информации, но также являются важным шагом к созданию баз знаний, порождающих новые знания или гипотезы для дальнейших исследований с помощью процедур автоматического вывода [Варфоломеев и Кравцов, 2007]. Следует отметить, что представленные в литературе модели семантических публикаций фокусируются, как правило, на иерархическую репрезентацию метаданных о текстах в семантических слоях, которые «накладываются» на электронную публикацию исследовательских текстов или текстов исторических источников [Shotton et al., 2009; de Waard, 2010].

В течение уже ряда лет авторы настоящей статьи изучают возможности и перспективы представления информации древнерусских исторических источников в семантических публикациях [см., например: Варфоломеев и Иванов, 2010, 2011б; Иванов и Варфоломеев, 2009, 2012; Иванов, 2012]. Не вызывает сомнения, что такие публикации должны удовлетворять всем общепринятым требованиям археографии, которые определяют правила транскрипции текстов и репрезентации метаданных, а также состав, полноту и особенности научного аппарата изданий источников.

При этом электронные докумен-ты — транскрипции (желательно дипломатические) текстов источников — должны сопровождаться сканированными изображениями оригиналов для проверки и коррекции передачи текстов. Главная же цель семантических из-даний — создание возможностей для эффективного проведения внутренней и внешней критики исторических источников, а также исторического синтеза на основе выявленной и верифицированной источниковой информации. При этом основным условием успешного применения компьютерных технологий в источниковедческой критике становится, с одной стороны, формирование ре-презентативных выборок публикуемых источников, адекватно отражающих исторически сложившиеся их комплексы [см. Иванов, 2010], а с другой сторо-ны, — возможность семантического связывания информации в пределах та-ких комплексов.

Однако существующие и апробированные авторами статьи модели семантических публикаций создают широкие возможности для семантического связывания прежде всего метаинформации, т.е. генерированной исследова-телем внетекстовой информации об источниках — об их происхождении, бытовании в контексте времени их создания, последующем хранении, об их палеографических особенностях, их публикациях и использовании в науч-ных трудах и т.п. К сожалению, имеющиеся модели семантических публика-ций не позволяют непосредственно представить в семантических сетях сви-детельства (информацию) источников: семантическая репрезентация тексто-вой источниковой информации предполагает их предварительную разметку, выявление логической последовательности структурных частей, увязывание формальной и семантической структуры текстов [Иванов и Варфоломеев, 2005; Ivanovs and Varfolomeyev, 2005; Варфоломеев и Иванов, 2011a].

В ходе разработки полифункционального прототипа семантической публикации древнерусских грамот XIII века, хранящихся в Латвийском госу-дарственном историческом архиве — структурном подразделении Латвий-ского национального архива и входящих в состав обширного комплекса до-кументальных источников «Moscovitica–Ruthenica» [Иванов, 2004], авторами были использованы инструменты семантических Wiki-систем, в частности, Semantic MediaWiki [SMW] со специальными средствами упрощенной раз-метки для выделения в тексте структурных частей, терминов и имен, а также добавления к тексту метаинформации [Ivanovs and Varfolomeyev, 2011]. Дан-ная разметка позволяет представить в семантических сетях информацию, от-ражающую лишь наиболее существенные аспекты содержания исторических источников. Альтернативная разметка текстов на основе схем разметки TEI и CEI [TEI; CEI] является более гибкой, но и она не способна исчерпывающе представить в семантической публикации содержание источников. К тому же разметка текстов, в сущности, опирается на определенную, весьма субъек-тивную интерпретацию содержания источников. Таким образом, полноценная семантическая публикация не может огра-ничиваться лишь репрезентацией источниковой информации, которая выде-лена исследователем в ходе разметки текстов документов. Необходимо так-же непосредственное представление в семантической сети смысла текстов исторических источников, для чего можно использовать один из так назы-ваемых «контролируемых естественных языков» («controlled natural languages») [Controlled Natural Language, 2010]. Тексты исторических источни-ков на таком языке переводятся компьютером в формулы логики предикатов первого порядка, над которыми можно производить логический вывод, по-лучая новые знания.

В настоящее время контролируемые естественные языки, служащие для представления знаний, стремительно развиваются. Для этих языков характерна жестко ограниченная семантика и упрощенный синтаксис. Можно назвать следующие контролируемые естественные языки: Controlled English to Logic Translation — CELT [Pease and Murray, 2003], Processable English — PENG [Schwitter, 2011], Computer-Processable Language — CPL [Clark et al., 2005], а также Attempto Controlled English — ACE [de Coi et al., 2009; Kuhn, 2009]. Attempto Controlled English — один из самых выразительных и распространенных в этом ряду, поэтому данный язык и был выбран для репрезентации содержания источников в прототипе семантической публикации.

Тексты на ACE могут быть переведены в DRS (Discourse Representation Structures) и, соответственно, напрямую трансформированы в формулы логики предикатов первого порядка [Kamp and Reyle, 1993]. Здесь следует отметить, что тексты на обычных языках не могут быть напрямую (автоматически) преобразованы в логические формулы вследствие неопределенности любого дискурса, поэтому смысл таких текстов может быть понят только при наличии определенного контекста. ACE выглядит совершенно естественным языком, хотя, на самом деле, это формальный язык, а именно — язык логики предикатов первого порядка, использующий синтаксис английского языка. Этот язык понятен и человеку, и компьютеру. В настоящее время Attempto Controlled English поддерживается рядом специализированных инструментов [Attempto Project]: синтаксический анализатор Attempto Parsing Engine (APE), который переводит тексты на ACE в DRS; модуль Attempto Reasoner (RACE), служащий для осуществления операций логического вывода на основе утверждений, записанных на ACE; редактор для корректировки текстов на ACE и др. Существенно и то, что для записи текстов на ACE могут использоваться различные специализированные словари. К тому же тексты исторических источников на ACE, обработанные при помощи модуля RACE, могут использоваться для генерирования новых гипотез на основе фактов, выявленных исследователем. Таким образом, Attempto Controlled English полностью отвечает задачам семантических публикаций исторических источников.

При этом неизбежно возникает вопрос о том, насколько приемлемо использование естественного контролируемого языка ACE для создания семантической публикации исторических источников на древнерусском языке, ведь любой перевод оригинального текста на другой язык (это относится и к передаче древнерусских документов на современном русском языке) неизбежно опирается на интерпретацию содержания источника, что ведет к частичному изменению смысла текста, а так же к частичной утрате его содержательных аспектов, не говоря уже о стилистических и выразительных средствах выражения содержания, которые не могут быть сохранены в переводе. При этом передача даже англоязычных текстов на ACE не может избежать существенных упрощений вследствие жестких правил перевода текстов.

В этой связи можно отметить, что, во-первых, передача оригинальных текстов на языке ACE служит лишь для создания возможности многоаспектного семантического связывания информации ряда источников, включенных в семантическую сеть и ни в коем случае не «замещает» оригинальный текст источника «суррогатным» текстом, поскольку тексты на ACE в семантической публикации должны быть соотнесены (связаны) с текстами на языке оригинала, представленными в рамках той же семантической сети.

То есть ACE является всего лишь инструментом генерирования семантической сети. Во-вторых, тексты на ACE позволяют включить в систему семантических связей значительно больший объем источниковой информации нежели схемы разметки SMW, TEI или CEI. В принципе, тексты источников, переданные на языке ACE, практически без потерь включаются в семантическую сеть, что значительно повышает эвристическую эффективность публикации. Для оценки применимости Attempto Controlled English для создания семантической публикации древнерусских грамот далее приводится текст одной из грамот, включенных в прототип семантической публикации, переданный на ACE.

Оригинальный текст	Перевод на язык АСЕ
† бл҃гблнне · ѿ мнтрофолнта рнзѣкого · ко своему мномоу · сн҃кн князю вєанкому феѡдору · н кѣ его дѣтемь · н ко владѹѣ н къ намѣстѣннкѹ · н ко всемъ коѡромъ ·	The metropolitan of Riga blesses the prince of Smolensk, the children of the prince of Smolensk, the bishop of Smolensk, the governor of Smolensk, and the boyars.
то вѹ дн тобе вєдомо · про тѹю жалобѹ · что внтѣбл ^а нне жаловѣанса на рнжанѣ · чмь то хотѣ ^а ѡправкнтиса протнву гѣамнка ·	The metropolitan of Riga demands that the prince of Smolensk knows-about a com- plaint that is lodged against some inhabi- tants of Riga by some inhabitants of Vitebsk. The Metropolitan of Riga sup- poses that the inhabitants of Vitebsk want to justify their actions to Helmich.
н нхѣ сло та ково · хотєан са темь словомь ѡправкнті · н рєкан такъ предъ княземь браньскѣмь · вѣ рєхалн · н · тѣ мужь нзъ рнгѣ · н оубнлн ѹ ловека · н оузалн · і · тѣ вѣркоѣсковъ коск ^ы	The inhabitants of Vitebsk appeal-to the Prince of Briansk. The inhabitants of Vitebsk say that 50 Rigans kill a man and grab 10 berkovets of wax.
н нчине а мнтрофолнтѣ тако молкѹ · как то внтѣблнне неправдою · жаловѣанса · на рнж ^а нѣ ·	The metropolitan of Riga says that the in- habitants of Vitebsk complain-on the in- habitants of Riga unjustly.

и нѣне то естъ нѣ кѣдомо • аже рнжане сѹтъ в томѣ невинокати •	The metropolitan of Riga knows that the inhabitants of Riga are not guilty.
и нѣне а тому • днѣлюса • аже тѣон намѣстъннѣ слышадеть ксакого уаѣка • саоса •	The metropolitan of Riga is surprised by the credulity of the governor of Smolensk.
а та правда естъ промѣжн касъ и насъ • кдѣ са тажа поунеть тѹ ко нцати	There is an agreement that is concluded by Riga and Smolensk. The agreement states that if a controversy arises then it must be settled by Riga and Smolensk.
и нѣне а молюса камъ • как то моѹнте стоати оу тои правдѣ • н оу крѣтномѣ уелоканнн •	The metropolitan of Riga demands that the agreement is observed.
аже иметь жалобнѣнса касъ кто на рнжанѣ • нан гѣамнко нан кто нѣн • н бѣ шантѣ к намъ • а мѣ правдоу да по бѣн правдѣ •	The agreement settles that if anyone com- plains-about Rigans, or Helmich, or somebody else then the prince of Smolensk must send a missive to the metropolitan of Riga. Then the metropolitan of Riga renders justice in accordance with divine justice.

Нельзя также не отметить, что и информация, относящаяся к источнику (факты об источнике) может быть переведена на ACE: «The charter_6 is written by the archbishop of Riga»; «The charter_6 probably refers to the charter_4»; «The charter_6 mentions Helmich and the prince of Briansk» и т.п. К тому же ACE позволяет записывать аксиомы и правила логического вывода, например: «if something X is written by someone Y, then Y is the author of X». В результате большой объем фактического материала, относящегося к текстам источников, который вводится и обрабатывается вручную в ходе создания семантической публикации на основе, например, технологии Semantic MediaWiki, может быть введен и обработан автоматически благодаря возможностям, которые предоставляет ACE. При этом, как уже отмечалось, но-вые знания — факты и гипотезы — об источниках и их текстах в перспективе могут быть получены на основе автоматического вывода. Поэтому язык ACE можно считать многообещающим инструментом для производства семанти-ческих публикаций комплексов исторических источников.

Работа выполняется при финансовой поддержке Программы стратегического развития ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности.

Список литературы

Варфоломеев и Иванов, 2010 — Варфоломеев А.Г., Иванов А.С. Семантические публикации комплексов исторических источников // Информационные технологии и письменное наследие. El'Manuscript-10: Материалы международной научной конференции (Уфа, 28–31 октября 2010 г.). Уфа; Ижевск: Вагант, 2010. С. 42–46.

Варфоломеев и Иванов, 2011a — Варфоломеев А.Г., Иванов А.С. Модели структуры и содержания исторических источников // Информационный бюллетень Ассоциации «История и компьютер». № 37. Специальный выпуск: Труды международной конференции «Компьютерные технологии и математические методы в исторических исследованиях (Петрозаводск, 11–16 июля 2011 г.). Петрозаводск, 2011. С. 25–31.

Варфоломеев и Иванов, 2011б — Варфоломеев А.Г., Иванов А.С. Семантические публикации информации исторических источников на основе технологии Wiki // *Vēsture: Avoti un cilvēki. Humanitārās fakultātes XX starptautisko zinātnisko lasījumu materiāli. Vēsture XIV. Daugavpils: Saule, 2011. С. 339–347.*

Варфоломеев и Кравцов, 2007 — Варфоломеев А.Г., Кравцов И.В. Приобретение и представление знаний в сетевом сообществе исследователей текстов // Вторая конференция «Системный анализ и информационные технологии» САИТ-2007, Обнинск, 10–14 сентября 2007 г.: Тр. конф.: в 2 т. Т. 1. Москва, 2007. С. 104–106.

Иванов, 2004 — Иванов А.С. «Moscowitica–Ruthenica» в Латвийском государственном историческом архиве: история формирования комплекса, состав и введение в научный оборот // *Древняя Русь: Вопросы медиевистики. 2004. № 3 (17), сентябрь. С. 47–54; № 4 (18), декабрь. С. 94–106.*

Иванов, 2010 — Иванов А.С. Источниковедческие и археографические аспекты реконструкции исторических комплексов источников (отдел «Moscowitica–Ruthenica» в бывшем архиве Рижского магистрата) // *Сословия, институты и государственная власть в России. (Средние века и раннее Новое время): Сборник статей памяти академика Л.В.Черепнина. Москва: Языки славянских культур, 2010. (Studia philologica). С. 97–105.*

Иванов, 2012 — Иванов А. Археография и актовое источниковедение в эпоху компьютерных технологий // *Проблемы дипломатики, кодикологии и актовой археографии: Материалы XXIV Международной научной конференции. Москва 2–3 февраля 2012. Москва: РГГУ, 2012. С. 321–323.*

Иванов и Варфоломеев, 2005 — Иванов А.С., Варфоломеев А.Г. Технология XML как инструмент компьютерного источниковедения (на примере формулярного анализа документов приказного делопроизводства) // *Круг идей: Алгоритмы и технологии исторической информатики: Труды IX конференции Ассоциации «История и компьютер». Москва; Барнаул, 2005. С. 241–281.*

Иванов и Варфоломеев, 2009 — Иванов А.С., Варфоломеев А.Г. Публикация и анализ рукописных исторических документов с помощью технологии XML // *Humanitāro Zinātņu Vēstnesis. 2009. № 16. С. 69–84.*

Иванов и Варфоломеев, 2012 — Иванов А.С., Варфоломеев А.Г. Компьютерный анализ структуры исторических источников (на примере отписок XVII в.) // *Vēsture: Avoti un cilvēki: Humanitārās fakultātes XXI starptautisko zinātnisko lasījumu materiāli. Vēsture XV. Daugavpils: Saule, 2012. С. 112–121.*

Иванов и Кузнецов, 2009 — Иванов А., Кузнецов А. Смоленско-рижские акты: XIII в. – первая половина XIV в.: Документы комплекса Moscowitica–Ruthenica об отношениях Смоленска и Риги. Рига, 2009.

Ahonen and Hyvönen, 2009 — Ahonen E., Hyvönen E. Publishing Historical Texts on the Semantic Web — A Case Study // *Proceedings of the Third IEEE International Conference on Semantic Computing (ICSC2009). Berkeley, 2009. Pp. 167–173.*

Attempto Project — Attempto Project Tools. <http://attempto.ifi.uzh.ch/site/tools/>. Baruzzo et al., 2009 — Baruzzo A. et al. Toward Semantic Digital Libraries:

Exploiting Web2.0 and Semantic Services in Cultural Heritage // *Journal of Digital Information. 2009. Vol. 10. No. 6. <http://journals.tdl.org/jodi/article/viewArticle/688/576>.*

CEI — Charters Encoding Initiative. <http://www.cei.lmu.de>.

Clark et al., 2005 — Clark P. et al. Acquiring and Using World Knowledge Using a Restricted Subset of English // The 18th International FLAIRS Conference FLAIRS'05. 2005. <http://www.cs.utexas.edu/users/pclark/papers/flairs.pdf>.

de Coi et al., 2009 — de Coi J. L. et al. Controlled English for Reasoning on the Semantic Web // Semantic Techniques for the Web. Lecture Notes in Computer Science, 5500. 2009. Pp. 276–308.

Controlled Natural Language, 2010 — Controlled Natural Language. Workshop on Controlled Natural Language, CNL 2009, Marettimo Island, Italy, June 8– 10, 2009. Revised Papers // Lecture Notes in Computer Science, 5972. 2010.

Ivanovs and Varfolomeyev, 2005 — Ivanovs A., Varfolomeyev A. Editing and Exploratory Analysis of Medieval Documents by Means of XML Technologies // Humanities, Computers and Cultural Heritage: Proceedings of the XVIth International Conference of the Association for History and Computing. Amsterdam, 2005. Pp. 155–160.

Ivanovs and Varfolomeyev, 2011 — Ivanovs A., Varfolomeyev A. Semantic Publications of Charter Corpora (The Case of a Diplomatic Edition of the Complex of Old Russian Charters ‘Moscowitica–Ruthenica’) // International Conference “Digital Diplomats: Tools for the Digital Diplomatist”. Napoli, 29 settembre – 1 ottobre 2011. Naples, 2011. Pp. 36–40.

Kamp and Reyle, 1993 — Kamp H., Reyle U. From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Dordrecht; Boston; London: Kluwer Academic Publishers, 1993.

Kuhn, 2009 — Kuhn T. How Controlled English can Improve Semantic Wikis // Proceedings of the Fourth Workshop on Semantic Wikis: European Semantic Web Conference 2009. [S.l.], 2009. Pp. 1–16 (SEUR Workshop Proceedings, vol. 464).

Mirzaee et al., 2005 — Mirzaee V., Iverson L., Hamidzadeh B. Computational Representation of Semantics in Historical Documents // Humanities, Computers and Cultural Heritage: Proceedings of the XVIth International Conference of the Association for History and Computing. Amsterdam, 2005. Pp. 199–206.

Pease and Murray, 2003 — Pease A., Murray W. An English to Logic Translator for Ontology-Based Knowledge Representation Languages // Proceedings of the 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China. [S. l.], 2003. Pp. 777–783.

Schwitter, 2011 — Schwitter R. Processing Coordinated Structures in PENG Light // AI 2011: Advances in Artificial Intelligence. Lecture Notes in Computer Science, 7106. 2011. Pp. 658–667.

Shotton, 2009 — Shotton D. Semantic Publishing: The Coming Revolution in Scientific Journal Publishing // Learned Publishing. 2009. Vol. 22. No. 2. Pp. 85–94.

Shotton et al., 2009 — Shotton D. et al. Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article // PLoS Computational Biology. 2009. Vol. 5. No. 4.
e1000361 <<http://dx.doi.org/10.1371/journal.pcbi.1000361>.

SMW — Semantic MediaWiki. <http://www.semantic-mediawiki.org>.

TEI — Text Encoding Initiative. <http://www.tei-c.org>.

de Waard, 2010 — de Waard A. From Proteins to Fairytales: Directions in Semantic Publishing // IEEE Intelligent Systems. 2010. Vol. 25. Issue 2. Pp. 83–88.