

Оригинальная статья / Original Article

DOI: 10.31857/S160578800025499-4

О создании русскоязычной онтологии по цифровой гуманитаристике

© 2023 г. А. Б. Антопольский

Доктор технических наук,
главный научный сотрудник Института научной информации
по общественным наукам РАН (ИНИОН РАН),
Россия, 117997, Москва, Нахимовский проспект, д. 51/21
ale5695@yandex.ru

Резюме. Предлагается разработка русскоязычной онтологии по цифровой гуманитаристике как части инфраструктуры этой дисциплины, которая получает организационное оформление как в мире, так и в России. Описываются основные зарубежные источники для такой онтологии. Также описываются российские проекты последних лет в области понятийного состава лингвистики и филологии в целом, которые целесообразно использовать. Кратко излагается формирование концептуальной модели онтологии. В качестве технологической основы предлагается модель проекта Ontolex-Lemon на платформе Семантической сети и лингвистических связанных открытых данных. Подчеркивается необходимость коллаборации различных специалистов и организаций.

Ключевые слова: цифровая гуманитаристика, онтологии, мировой опыт, российские источники, компьютерная лингвистика, цифровая филология, Ontolex-Lemon.

Для цитирования: Антопольский А.Б. О создании русскоязычной онтологии по цифровой гуманитаристике // Известия Российской академии наук. Серия литературы и языка. 2023. Т. 82. № 2. С. 22–32. DOI: 10.31857/S160578800025499-4

About the Creation of a Russian-Language Ontology for Digital Humanities

© 2023 Alexander B. Antopolsky

Doct. Sci. (Tech.),
Head Researcher at the Institute of Scientific Information
for Social Sciences of the RAS (INION RAN),
51-21 Nakhimovskiy Prospect, Moscow, 117997, Russia
ale5695@yandex.ru

Abstract. It is proposed to develop a Russian-language ontology for digital humanities as part of the infrastructure of this discipline, which receives organizational registration both in the world and in Russia. The main foreign sources for such an ontology are described. It also describes the Russian projects of recent years in the field of conceptual composition of linguistics and philology in general, which are advisable to use. The formation of the conceptual model of ontology is briefly described. As a technological basis, the Ontolex-Lemon project model is proposed on the platform of a Semantic Network and linguistic linked open data. The necessity of collaboration of various specialists and organizations is emphasized.

Key words: digital humanities, ontologies, world experience, Russian sources, computational linguistics, digital philology, Ontolex-Lemon.

For citation: Antopolsky, A.B. *O sozdanii russkoyazychnoj ontologii po tsifrovoj gumanitaristike* [About the Creation of a Russian-Language Ontology for Digital Humanities]. *Izvestiâ Rossijskoj akademii nauk. Seria literatury i âzyka* [Bulletin of the Russian Academy of Sciences: Studies in Literature and Language]. 2023, Vol. 82, No. 2, pp. 22–32. (In Russ.) DOI: 10.31857/S160578800025499-4

Введение

Можно утверждать, что с начала XXI в. в мире сложилась новая дисциплина, которую принято называть Digital Humanities, или цифровой гуманитаристикой. В монографии [1], а также в исследовании, описанном в работе [2], показано, что в мире уже сотни организаций относят себя к этой дисциплине, активно реализуются международные и национальные программы в этой области и, самое главное, создается специализированная инфраструктура, поддерживающая исследования, разработки и формирование информационного пространства (инфосферы) цифровой гуманитаристики. Описанию состава и функционирования этой инфраструктуры в Евросоюзе посвящена другая публикация автора [3].

Нужно отметить, что состав и содержание цифровой гуманитаристики остаются дискуссионными. В частности, различные подходы к определению данной дисциплины представлены в хрестоматии, изданной в Красноярске [4]. Однако большинство исследователей сходятся во мнении, что ядром цифровой гуманитаристики являются такие дисциплины, как историческая информатика, компьютерная лингвистика, цифровая филология, а также применение информационно-компьютерных технологий в библиотеках, архивах и музеях. Из этого, в частности, следует, что цифровая гуманитаристика должна стать одним из важных и перспективных направлений деятельности Отделения историко-филологических наук РАН. Действительно, одна из последних научных сессий ОИФН РАН была посвящена этим разработкам [5].

В России цифровая гуманитаристика также получает организационное оформление: в нескольких университетах созданы соответствующие структуры, которые объединились в Российскую ассоциацию цифровых гуманитарных наук¹. Также ведется работа по созданию инфраструктуры: ИНИОН РАН и Центр цифровых гуманитарных наук Сибирского федерального университета начали проект по созданию информационно-справочной системы по цифровой гуманитаристике [6].

Важным компонентом инфраструктуры современной цифровой науки являются онтологии, обеспечивающие представление понятийного и лексического состава различных научных дисциплин, а также научного знания в целом. Существенно, что онтологии отражают

концептуальные модели как в формализованном виде, пригодном для компьютерной обработки, так и в естественно-языковой форме, учитывающей специфику национальной терминологии. Эти онтологии должны применяться при реализации различных проектов, создавая логическую и понятийную основу представления научной информации. При этом, кроме своей основной функции, онтология одновременно должна организовывать пространство метаданных, которое описывает, представляет и структурирует инфосферу данной области знания.

В данной статье обосновывается возможность и целесообразность создания русскоязычной онтологии по цифровой гуманитаристике с учетом как мирового опыта, так и отечественных разработок, которые можно положить в основу создания онтологии.

Мировые источники онтологии

Очевидно, что цифровая гуманитаристика носит комплексный междисциплинарный характер, поэтому средства индексирования для этой сферы должны интегрировать понятия из разных областей. Европейские исследователи коллективными усилиями разработали специальную таксономию для представления понятий по исследованиям в цифровой гуманитаристике, которая получила название TADIRAN². Как пишут разработчики, “эта таксономия цифровой исследовательской деятельности в гуманитарных науках была разработана для использования сайтами и проектами, управляемыми сообществом, которые направлены на структурирование информации, относящейся к цифровым гуманитарным наукам, и облегчение ее обнаружения. Ожидается, что таксономия будет особенно полезна для усилий, направленных на сбор информации о цифровых гуманитарных инструментах, методах, проектах”.

TADIRAN включает 3 фасета:

- Виды исследовательской деятельности
- Объекты исследования
- Методы исследований

TADIRAN представляет собой относительно небольшой инструмент. *Виды деятельности* включают 8 рубрик и 40 подрубрик, *Объекты* — 40 наименований типов в виде плоского списка, и *Методы* включают 41 термин, часть из которых

¹ Российская ассоциация цифровых гуманитарных наук <http://dhrussia.ru/> (дата обращения: 01.04.2022).

² TaDiRAH — Taxonomy of Digital Research Activities in the Humanities. URL: <http://tadirah.dariah.eu/vocab/index.php> (дата обращения: 01.04.2022).

привязана к *Видам деятельности*, а часть объявлена независимыми.

Наименования рубрик и подразбук *Видов деятельности* снабжены достаточно подробными толкованиями. *Объекты* и *Методы* толкований не имеют.

Анализ TADIRAN показывает, что предлагаемый в нем состав понятий и их связей лишь частично отражает российскую терминологическую традицию и практику применения понятий. Это видно с первой же рубрики основного фасета *Виды деятельности*, которая называется *Захват* (в оригинале — capture).

Приведем толкование этого понятия по TADIRAN: “Захват обычно относится к деятельности по созданию цифровых суррогатов существующих культурных артефактов или выражению существующих артефактов в цифровом представлении (оцифровка). Это может быть ручной процесс (как в расшифровке) или автоматизированная процедура (как в визуализации или распознавании данных). Такой захват предшествует обогащению и анализу, по крайней мере, с систематической точки зрения, если не на практике”.

Понятие *Захват* включает такие подразбук, как *Преобразование*, *Распознавание данных*, *Раскрытие*, *Сбор*, *Отображение*, *Запись*, *Транскрипция*, а в качестве метода указан только *Обход веб-страниц*.

Предлагаемое толкование так же, как и содержание рубрики, вызывает массу вопросов. Начнем с того, что понятия *Захват* в российской традиции вообще нет. Вероятно, ближе всего к содержанию этого понятия относится *Сбор*. Но тогда *Преобразование*, *Распознавание данных*, *Раскрытие*, *Транскрипция* должны относиться к виду деятельности *Обработка информации*, но не к сбору.

А создание цифровых суррогатов — это, конечно, прежде всего, *Оцифровка*. Однако этот метод в TADIRAN отнесен не к *Захвату*, а к *Созданию*.

Различные вопросы возникают при анализе и других видов деятельности и методов, включенных в TADIRAN. В целом эта система понятий не вполне соответствует российской практике.

Более привычен для российского исследователя перечень типов объектов TADIRAN. Однако и здесь есть вопросы. Так, например, сложно различать такие типы объектов, как *Проекты*, *Исследования*, *Процесс исследования*, *Результаты исследований*. Неясно соотношение понятий

Программное обеспечение и *Инструменты*. К тому же в этом перечне в едином ряду представлены понятия самых разных семантических категорий, от тематических рубрик до абстрактных понятий.

Добавим, что в TADIRAN отсутствует множество понятий, необходимых для отображения понятий цифровой гуманитаристики, например, тематические понятия, языковой, географический, хронологический и другие фасеты.

Таким образом, хотя TADIRAN, конечно, должна быть учтена при разработке отечественной онтологии по цифровой гуманитаристике, но не может быть ее основой.

Конечно, за рубежом разработано и используется множество других средств представления систем понятий в цифровой гуманитаристике. Эти средства реализованы в виде стандартов, систем метаданных, тезаурусов, классификаций и собственно онтологий по разным областям знания и формам деятельности, которые имеют отношение к данной дисциплине.

Анализ и обобщение мирового опыта по представлению понятийных и лексических собраний, как универсальных, так и тематических, связанных с областью культурного наследия, сделаны Дж. Рилей из Университета МакГилла [7]. В секторе культурного наследия существует огромное количество стандартов метаданных и аналогичных понятийных схем, а их взаимосвязь еще больше осложняет ситуацию. В проекте Дж. Рилей собрано и систематизировано свыше 100 таких средств. Они загружены в платформу Dataverse, что позволяет удобно анализировать эти средства по различным основаниям, а также визуализировать результаты анализа.

В работе Дж. Рилей различные средства представления понятий рассматриваются как пространство метаданных, включающее разные стандарты. Каждый из 105 стандартов оценивается по степени его применения к определенным категориям по каждой из четырех осей: сообщество, область, функция и цель.

Под сообществами имеются в виду следующие категории: архивы, музеи, библиотеки, информационная индустрия.

Области включают такие категории: культурные объекты, наборы данных, геопространственные данные, движущиеся объекты, музыкальные материалы, научные данные, визуальные ресурсы.

Функциональные категории: концептуальные модели, стандарты на контент, контролируемые

словари, схемы и технологии, языки разметки, форматы, стандарты на структуру.

Категории целей предлагаются следующие: данные, описательные метаданные, метаданные оформления, метаданные сохранности, правовые метаданные, структурные метаданные, технические метаданные

Значение стандарта в данной категории определяется сочетанием его распространенности в этой категории, его конструктивного замысла и его общей пригодности для использования в этой категории.

Представленные в данной работе стандарты относятся к числу наиболее широко используемых или публикуемых в сообществе культурного наследия, хотя, конечно, учтены не все стандарты, которые могут иметь отношение к области культурного наследия.

Еще одна обобщающая работа, которую можно использовать в качестве источника онтологии цифровой гуманитаристики, — это информационная система связанных открытых словарей LOV³.

В LOV в настоящее время загружено почти 800 словарей, онтологий и других лексических и понятийных собраний, в том числе разнообразные стандарты метаданных.

Данные в LOV представлены в формате связанных открытых данных. Для каждого словаря в LOV термины (классы, свойства, типы данных, экземпляры) индексируются и предлагается функция полнотекстового поиска. По сравнению с другими существующими поисковыми системами онтологии алгоритм ранжирования в поисковой системе LOV не только основан на популярности термина в наборах данных, но и учитывает его популярность в экосистеме LOV. В системе возможен поиск словарных данных по семантическим категориям, таким как *Метаданные*, *Методы*, *Сервисы*, *Каталоги*, *Персоны*, *Словари*, *Мультимедиа*, *События*, *Время* и др.

В мире, конечно, разрабатывается множество других лексико-семантических собраний, которые могут быть использованы при разработке русскоязычной онтологии по цифровой гуманитаристике. Многие из них рассмотрены в обзорной монографии автора [8]. Однако наиболее актуален для данной задачи российский опыт.

Российский опыт

В данном разделе мы кратко представим российские источники словарно-энциклопедических

сведений, а также опишем практические реализованные концептуальные модели российских разработчиков последних лет. Тематически мы ограничимся областью компьютерной лингвистики и цифровой филологии.

В настоящее время российским пользователям Интернета доступны сотни как оцифрованных традиционных словарей на различных языках, так и специализированных лексикографических баз данных различного назначения. Известно несколько российских агрегаторов лексикографических ЛИР, назовем наиболее популярные: *Грамота.ру*⁴, *Словари*⁵, *Lingvo*⁶, *Словари Онлайн*⁷, *Мультитран*⁸, *Академик*⁹, *Словари.СС*¹⁰.

Специализированный подбор словарей и энциклопедий филологической тематики собран в соответствующем разделе известной информационной библиотеки ФЭБ “Русская литература и фольклор”¹¹.

Вероятно, наиболее авторитетным собранием компьютерных словарей в России является раздел “Лексикографические ресурсы” на сайте Института русского языка им. В.В. Виноградова РАН¹².

Сведения о российских лексикографических ЛИР для научных исследований собраны в Навигаторе информационных ресурсов по языкознанию¹³. Всего там описано около 100 лексикографических ЛИР, разработанных в учреждениях РАН, в том числе различные словари на основе НКРЯ.

⁴ Грамота.ру. URL: <http://gramota.ru/slovari/> (дата обращения: 01.04.2022).

⁵ Словари. URL: <http://slovari.ru/start.aspx?s=0&p=3050> (дата обращения: 01.04.2022).

⁶ Lingvo. URL: <https://www.lingvolive.com/ru-ru> (дата обращения: 01.04.2022).

⁷ Словари Онлайн. URL: <https://slovaronline.com/> (дата обращения: 01.04.2022).

⁸ Мультитран. URL: <https://www.multitran.com/> (дата обращения: 01.04.2022).

⁹ Академик. URL: <https://dic.academic.ru/> (дата обращения: 01.04.2022).

¹⁰ Словари. СС. URL: <https://slovar.cc/> (дата обращения: 01.04.2022).

¹¹ Фундаментальная электронная библиотека “Русская литература и фольклор”. Словари, энциклопедии. URL: <http://feb-web.ru/feb/feb/dict.htm> (дата обращения: 01.04.2022).

¹² Онлайн ресурсы ИРЯ РАН. Лексикографические ресурсы. URL: <https://www.ruslang.ru/onlayn-resursy-iry-a-ran> (дата обращения: 01.04.2022).

¹³ Навигатор информационных ресурсов по языкознанию. URL: <http://nirya.znion.ru/> (дата обращения: 01.04.2022).

³ LOV — Linked Open Vocabularies. URL: <https://lov.linkeddata.es/dataset/lov> (дата обращения: 01.04.2022).

Перечень российских электронных словарей и лексических баз данных можно найти в каталоге NL-Pub в разделах Словари и Тезаурусы¹⁴.

Российский опыт в области создания онтологий и компьютерной лексикографии в целом также достаточно богат и разнообразен. Обобщение этого опыта можно найти в том числе в монографиях [9] и [10], а также в работе автора [8].

В России наиболее полно и глубоко онтологическая проблематика реализована в Лаборатории информационных исследований¹⁵, которая известна своими разработками в области создания тезаурусов и автоматической обработки текста. Созданный в лаборатории тезаурус *RuThes*¹⁶ используется в информационной системе УИС Россия¹⁷ и в других проектах с государственными и коммерческими организациями.

Тезаурус *RuThes* представляет собой лингвистический ресурс концептуального типа, то есть иерархическую сеть понятий, к которым приписаны текстовые выражения. И в этом смысле *RuThes* относится к тому же классу, что и тезаурус *WordNet*. При этом, в отличие от *WordNet*, который создавался как модель человеческой памяти (раздельное описание частей речи, специальные типы отношений и др.), тезаурус *RuThes* задумывался именно как ресурс для автоматической обработки текстов.

Следует отметить также разработки компьютерных лингвистов из СПбГУ. Разработанный этим коллективом тезаурус *RussNet* стал одним из заметных российских лингвистических ресурсов¹⁸.

Важным достижением в области систематизации понятий и лексики в области компьютерной лингвистики стала разработка *Портала знаний*¹⁹.

В этом проекте подробно разработана классификация лингвистических ресурсов, методов

автоматического исследования текста, а также всей сопутствующей информации; фактически построена онтология понятий, относящихся к компьютерной лингвистике. Описание проекта можно найти в работе [10].

Приведем для иллюстрации фрагмент разработанной классификации.

Методы обработки текста

|| Методы анализа текста

||| Методы морфологического анализа

||| Методы разрешения анафоры

|| Методы разрешения неоднозначности

|||| Методы разрешения лексической многозначности

||||[⊥] Методы снятия грамматической неоднозначности

||| Методы сегментации текста

|||[⊥] Методы синтаксического анализа

||[⊥] Методы генерации текста

Оригинальным проектом по представлению лингвистических знаний является работа известного российского лингвиста С.А. Крылова. Результат этой работы, которую он назвал металингвистической БД, размещен на информационном портале *Starling*²⁰. Пр процитируем С.А. Крылова: «Металингвистические базы данных (МБД), служат инструментом систематизации знаний о лингвистике (а не напрямую о языке), однако косвенно способствуют также систематизации сведений о языке. Можно выделять две разновидности МБД:

(1) метанаучные (МН-) МБД (входы в которые являются металингвистическими проекциями научных текстов по лингвистике) и

(2) метаобъектные (МО-) МБД (входы в которые являются металингвистическими проекциями языковых сущностей).

Входами в МО-МБД служат, например, характеристики языковых общностей (лингвонимические, этнонимические, топонимические, хронологические); нарицательные лингвистические термины; имена языковых единиц (в том числе имена таксономических классов внеязыковых сущностей).

Следует прежде всего проводить различие между онтологическим (материальным) уровнем,

¹⁴ Ресурсы. Словари. URL: <https://nlpub.ru/Ресурсы#.D0.A1.D0.BB.D0.BE.D0.B2.D0.B0.D1.80.D1.8C> (дата обращения: 01.12.2021).

¹⁵ Лаборатория информационных исследований. URL: <http://www.labinform.ru/> (дата обращения: 01.04.2022).

¹⁶ О лингвистической онтологии «Тезаурус РуТез» см.: URL: <http://www.labinform.ru/pub/ruthes/index.htm> (дата обращения: 01.04.2022).

¹⁷ УИС Россия. URL: <https://www.uisrussia.msu.ru/> (дата обращения: 01.04.2022).

¹⁸ RussNet. URL: http://project.phil.spbu.ru/RussNet/index_ru.shtml (дата обращения: 01.04.2022).

¹⁹ Компьютерная лингвистика. Портал знаний. URL: <https://uniserv.iis.nsk.su/cl/> (дата обращения: 01.04.2022).

²⁰ Вавилонская башня. Проект «Эволюция языка». URL: <https://starling.rinet.ru/program.php?lan=ru> (дата обращения: 01.12.2021).

на котором можно выделить объектное множество (оригинал, универсум) с существующими в нём отношениями, и гносеологическим (эпистемологический, идеальный) уровнем, на котором выделяется модельное множество (модель, теория) с заданными на нём отношениями. Эту модель и строит металингвист, воплощающий её в виде грамматики, словаря, предметного или именного указателя, таблицы, графа, дерева, карты, атласа, базы данных и т.п.” [11].

В данной работе предлагается развернутая система понятий, представляющих предметную область; мы приводим верхние уровни этой классификации.

I. Универсум языковых явлений.

IA. Общелингвистический универсум.

(IA.1.) Мир языковой системы.

(IA.1.0.) Языковая система и ее подсистемы.

(IA.1.1.) Языковые единицы (ЯЕ).

(IA.1.2.) Отношения между ЯЕ.

(IA.1.3.) Члены отношений между ЯЕ.

(IA.1.4.) Функции ЯЕ.

(IA.1.5.) Способы выражения значений.

(IA.1.6.) Классы ЯЕ.

(IA.1.7.) Члены классов ЯЕ.

(IA.1.8.) Языковые структуры.

(IA.1.9.) Части языковых структур.

(IA.1.10.) Языковые процессы.

(IA.1.11.) Логические связи языковых явлений.

(IA.2.) Речевая динамика.

(IA.3.) Речевая способность (типы, аспекты и компоненты).

(IA.4.) Речевое варьирование (типы и проявления).

(IA.5.) Языковое функционирование (типы).

(IA.6.) Языковые изменения (типы, аспекты и компоненты).

(IA.7.) Языковые сходства и различия (типы).

(IA.8.) Исторические отношения между языковыми общностями.

IB. Частнолингвистический универсум.

(IB.1.) Универсум исторических языковых общностей.

(IB.2.) Универсум ареалов распространения языков: континенты, регионы, страны, населённые пункты.

(IB.3.) Универсум частнолингвистических единиц.

IV. Универсум речевых событий

(IV.1.) Универсум словесности (множество текстов).

(IV.1.2.) Универсум памятников письменности.

(IV.1.3.) Универсум высказываний.

(IV.2.) Универсум вхождений речевых знаков-экземпляров (tokens).

II. Универсум собственно лингвистики.

(II.1.) Лингвисты (в том числе лингвисты-непрофессионалы).

(II.2.) Лингвистические школы и направления.

(II.3.) Лингвистические кружки, общества, ассоциации и т.п.

(II.4.) Места, где протекает деятельность лингвистов (континенты, страны, провинции, населённые пункты).

(II.5.) Учреждения, где протекает деятельность лингвистов.

(II.6.) Универсум лингвистических работ.

III. Мир лингвистических моделей.

(III.1.) Описания языков (словари, грамматики и т.п.).

(III.2.) Описания речевых отрезков: транскрипции, хрестоматии текстов, издания памятников, продукты транскрипции и транслитерации, переводы текстов, фонетические сонограммы, комментарии, глоссы, формальные представления текстов в виде морфологических и синтаксических “разборов”, синтаксических графов (в частности, деревьев зависимостей и составляющих), цепочки трансформационного вывода, толкования отдельных примеров и т.п.

(III.3.) Описания ЯЕ: словарные статьи, правила, законы, исключения к правилам и т.п.

Ценным источником онтологической и лексико-семантической информации является русский *Викисловарь (WIKT)*²¹. Словарь был открыт 1 мая 2004 года, и сейчас в нём содержится 1 114 852 статьи о словах, словообразовательных единицах и словосочетаниях более 500 языков мира. Русский *Викисловарь* является восьмым по величине викисловарем, состоящим из более чем 520 000 статей. Одна статья представляет собой лексическую

²¹ Русский Викисловарь. URL: https://ru.wiktionary.org/wiki/Заглавная_страница (дата обращения: 01.04.2022).

запись, написанную более чем 120 000 пользователей с 2004 года.

Предметом описания в *Викисловаре* являются все лексические единицы, зафиксированные во всех письменных языках и диалектах мира. Словник *Викисловаря* формируется из лексем, морфем и других словообразующих единиц, а также устойчивых словосочетаний этих языков, с использованием тех графических основ (систем письменности), которые традиционно применяются в соответствующих языках.

Викисловарь сочетает функции нескольких видов традиционных словарей, включая толковые, орфографические, орфоэпические, грамматические, переводные, фразеологические и этимологические словари, а также тезаурусы.

Можно указать также на проект разработки онтологии научного знания на основе интеграции классификационных систем и толковых словарей. Это проект, который был выполнен в 2014–2015 гг. по заданию Минобрнауки РФ при головной роли ВИНТИ и в котором принимала участие группа специалистов из разных организаций, включая автора.

Подробное описание проекта имеется в работе [12]. Здесь укажем, что практическим результатом данного проекта было создание базы данных Терминологические словари²². Источниками формирования терминологических словарей являются следующие ресурсы:

- энциклопедические словари (энциклопедии), имеющиеся в свободном доступе через Интернет;
- ключевые слова, приписанные к рубрикам согласно технологии библиографических баз данных ВИНТИ, БЕН РАН и НПБ им. К.Д. Ушинского;
- база данных стандартизированной терминологии Росстандарта;
- лексика наименований рубрик сопоставляемых классификационных систем;
- тезаурусы тематических рубрик НПБ им. К.Д. Ушинского и ИФП СО РАН.

Конечной целью проекта предполагалось создание онтологии научного знания на основе интеграции всех перечисленных источников. Однако проект был завершен формированием экспериментальной базы данных, где процессы интеграции были реализованы лишь для

нескольких тематических областей, включая языкознание.

Еще один проект, реализованный с участием автора, был направлен на исследование лексики метаданных российских лингвистических ресурсов. Целью его являлось создание лексической и терминологической основы для полноценной онтологии по лингвистике (языкознанию). Продукт, который получил название *Онтология поисковых терминов по лингвистике (ОПТЕЛ)*, также может служить для навигации и/или метапоиска в российских ЛИР. Принципы отбора источников для ОПТЕЛ, структура БД, особенности отдельных словарей метаданных описаны в работе [13]. В настоящее время ОПТЕЛ реализована и размещена в Интернете²³.

Реализованная версия ОПТЕЛ включает 55 словарей, использованных в 28 российских ЛИР разных типов. Всего в ОПТЕЛ представлено свыше 430 тыс. уникальных лексических единиц; объем каждого словаря, данные о пересечениях, а также другая статистика приведены в работе [14]. Все словарные статьи независимо от исходной структуры словаря метаданных приведены к классической тезауральной (дескрипторной) форме, как это описано в работе [15].

Функциональность ОПТЕЛ заключается в возможности исследовать лексику и парадигматику словарей информационных языков, вошедших в ОПТЕЛ, и, что особенно важно, сопоставлять словарные статьи для выработки общей (сводной) словарной статьи, устранения обнаруженных противоречий и исправления ошибок.

Из последних исследований, связанных с представлением системы понятий в области цифровой филологии, можно назвать Тезаурус по сравнительной поэтике и сравнительному литературоведению, реализованный в одноименной информационной системе²⁴.

Тезаурус состоит из терминов, находящихся в определенных отношениях. Система отношений представлена в виде “вертикальной” иерархической структуры, дополненной “горизонтальными” неиерархическими отношениями (синонимия, альтернативное положение в иерархии). Все термины разделяются на два типа: термин-дисциплина, определяющий некоторую область науки (переводоведение, поэтика, стиховедение и т.д.), и термин-объект, определяющий

²² Терминологические словари. URL: http://termin.jscc.ru/users/sign_in

²³ Онтология поисковых терминов по лингвистике. URL: <http://db.inion.ru/optel/> (дата обращения: 01.12.2021).

²⁴ Сравнительная поэтика и сравнительное литературоведение. URL: <http://cpcl.feb-web.ru/>

конкретное научное понятие (буквальный перевод, комедия, ямб и т.д.). Между терминами существуют следующие структурообразующие отношения: “рубрика — раздел” (между двумя терминами-дисциплинами, например: стиховедение — метрика), “дисциплина — объект дисциплины” (между термином-дисциплиной и термином-объектом, например: метрика — метр) и “род — вид” (между двумя терминами-объектами, например: метр — ямб).

Остальные отношения не являются структурообразующими: они эксплицируются в полях статей, но не отражаются в структуре (таковы, например, отношения между синонимами). Кроме того, в ряде случаев отношения между терминами носят принципиально сетевой характер. Для того чтобы сохранить “древесную” иерархию, в соответствующем месте структуры размещаются ссылки на альтернативное месторасположение термина (“термин-ссылка”).

В описанной системе имеется также раздел *Энциклопедия*, в котором содержатся краткие биографические и библиографические сведения о поэтах, переводчиках и исследователях-компаративистах.

Очевидно, что данный проект содержит все необходимые предпосылки для разработки онтологии, хотя и по ограниченной тематике.

Таким образом, можно констатировать наличие большого числа созданных российскими лингвистами и информатиками источников, которые могут быть использованы при создании онтологии по цифровой гуманитаристике, по крайней мере, ее филологического компонента.

В заключительной части статьи мы изложим некоторые предложения по методике разработки этой онтологии.

Принципы разработки модели онтологии

Международное сообщество не только ведет практическую работу по созданию онтологий и агрегации соответствующих данных, но и проводит исследования по разработке общей модели, предполагающей взаимодействие онтологий и классической лексикографии. Этой проблеме посвящен проект ONTOLEX²⁵, который реализует одноименное сообщество в рамках деятельности Консорциума Всемирной паутины. Основные задачи проекта:

- Разработать модели представления лексики (и машиночитаемых словарей) относительно

онтологий. Эти лексические модели предназначены для представления лексических записей, содержащих информацию о том, как элементы онтологии (классы, свойства, индивиды и т.д.) реализуются на нескольких языках.

- Продemonстрировать дополнительную ценность представления лексики в семантической сети, уделяя особое внимание тому, как использование принципов связанных данных может позволить повторно использовать существующую лингвистическую информацию из такого ресурса, как WordNet.
- Предоставить лучшие практики использования лингвистических категорий данных в сочетании с лексикой.
- Продemonстрировать, что создание такой лексики в сочетании с семантикой, содержащейся в онтологиях, может улучшить производительность инструментов автоматической обработки текста.
- Объединить людей, работающих над стандартами представления лингвистической информации (синтаксической, морфологической, семантической и прагматической), опираясь на существующие инициативы и определяя пути сотрудничества на будущее.
- Обеспечить взаимодействие между существующими моделями для представления и структурирования лингвистической информации.

Последний результат проекта — модуль лексикографии модели Lexicon для онтологий (lemon) как результат работы группы сообщества Ontology Lexicon (OntoLex). Модуль предназначен для представления словарей и любых других лингвистических ресурсов, содержащих лексикографические данные, и адресует структуры и аннотации, обычно встречающиеся в лексикографии. Этот модуль работает в сочетании с модулем lemon core, называемым OntoLex²⁶. Как нам представляется, именно модель OntoLex-Lemon должна быть положена в основу предлагаемой онтологии цифровой гуманитаристики.

Началом разработки любой онтологии предметной области должно стать формирование концептуальной модели E/R (Сущность / Отношение), определяющей основные процессы, субъекты и объекты в этой области и виды взаимодействий

²⁵ Ontology-Lexicon community group. URL: <https://www.w3.org/community/ontolex/> (дата обращения: 01.04.2022).

²⁶ The OntoLex Lemon Lexicography Module. URL: <https://www.w3.org/2019/09/lexicog/> (дата обращения: 01.04.2022).

между ними. Для разработки модели E/R бывает полезно сформулировать классические вопросы WHO, WHERE, WHEN и WHAT.

Для цифровой гуманитаристики основными объектами являются информационные ресурсы и инструментарий, субъектами — организации, физические лица, программы, процессами — сервисы, исследования, создание ресурсов и инструментов. Понятно, что процессы должны иметь назначение, хронологические, географические и инструментальные аспекты.

Рассмотренная нами в начале статьи таксономия TADIRAH содержит попытку формирования части концептуальной модели цифровой гуманитаристики, включающей цели исследований, методы исследований, объекты.

Исследование, проведенное с целью создания информационно-справочной системы по цифровой гуманитаристике и упомянутое в начале данной статьи, выявило основные сущности, определяющие состав инфосферы цифровой гуманитаристики: акторы, проекты, информационные ресурсы, инструменты, сервисы, нормативы.

Основными свойствами, характеризующими эти сущности и выделенными в ходе анализа, являются вид информационных ресурсов, назначение сервисов и инструментов, тематическая привязка ресурсов, географическая привязка актора, хронологическая привязка объекта.

Вероятно, необходимыми сущностями (или свойствами сущностей) для области цифровой гуманитаристики должны быть категории, выделенные в цитируемом выше исследовании Дж. Рилей: институции, к которым принадлежит объект (архив, музей, библиотека и др.), функциональность и цели стандартов метаданных.

После определения состава сущностей области цифровой гуманитаристики, классов их свойств и отношений можно переходить к формированию списков понятий каждой сущности и троек “субъект—отношение—объект”, “объект—отношение—свойство” и других для выявленных понятий. Для этой цели можно использовать существующие как зарубежные, так и российские энциклопедии, классификации, таксономии, онтологии, стандарты.

Заключительный шаг — это лексикализация понятий онтологии на основе существующих тезаурусов и словарей.

Как уже было отмечено, технологической основой, в рамках которой предлагается создавать онтологию цифровой гуманитаристики, должна

быть, вероятно, технология связанных открытых данных. Конкретная технологическая модель реализована в проекте Ontolex-Lemon, описанном выше.

В заключение несколько слов об организации деятельности по созданию онтологии. По нашему мнению, это должен быть коллаборативный проект, отражающий коллективное знание научного сообщества, представленный в сети и доступный для постоянного пополнения участниками проекта, однако с централизованной модерацией. В проекте должны иметь возможность принять участие специалисты различных гуманитарных дисциплин, а также компьютерные лингвисты и информационные технологи. Технологии коллаборативных проектов хорошо известны, существует открытое программное обеспечение для коллективного создания онтологий; дело за организацией и желанием. Вероятно, инициативу по созданию онтологии должны проявить структуры, входящие в Ассоциацию цифровых гуманитарных наук. Автор надеется, что в число участников предлагаемого проекта войдет также ИНИОН РАН и другие академические организации гуманитарного профиля.

СПИСОК ЛИТЕРАТУРЫ

1. Digital Humanities: гуманитарные науки в цифровую эпоху / под ред. Г.В. Можяевой. Томск: Изд-во Том. ун-та, 2016. 120 с.
2. Антопольский А.Б. Инфосфера цифровой гуманитаристики: опыт анализа // Информационные ресурсы России. 2022. № 1. С. 30–38. DOI: 10.52815/0204-3653_2022_01185_30
3. Антопольский А.Б. Информационная инфраструктура социально-гуманитарных наук в Евросоюзе // Наука и научная информация / Scholarly Research and Information. 2021. № 4(1). С. 18–32. DOI: 10.24108/2658-3143-2021-4-1-18-32
4. Цифровые гуманитарные науки / Хрестоматия. URL: <http://lib3.sfu-kras.ru/ft/LIB2/ELIB/b71/free/i-531505996.pdf> (дата обращения: 01.04.2022).
5. Гуманитарные науки в эпоху цифровизации. Видеотрансляция Общего собрания Отделения историко-филологических наук РАН. URL: <http://hist-phil.ru/events/427/> (дата обращения: 01.04.2022).
6. Антопольский А.Б. Инфосфера цифровой гуманитаристики: структура и тенденции развития // Электронное информационное пространство для науки, образования, культуры: материалы IX Международной научно-практической

- конференции (г. Орел, 16–17 декабря 2021 г.) / науч. ред. и сост. В.В. Матвеев, Д.Н. Грибков, ред. кол.: А.А. Аксехин, С.Н. Манько. Орел: Орловский гос. ин-т культуры, 2022. С. 5–11.
7. Riley J. Seeing Standards: A Visualization of the Metadata Universe. DOI: 10.5683/SP2/UOHPVH, Scholars Portal Dataverse, V3. URL: gWl/jicj8wtJm4Grmph7TQ== [fileUNF] (дата обращения: 01.04.2022).
 8. Антопольский А.Б. Лингвистические информационные ресурсы: монография / А.Б. Антопольский; ИНИОН РАН, Фундаментальная; науч. ред Д.В. Ефременко, М.: ИНИОН РАН, 2022. 466 с.
 9. Рубашкин В.Ш. Онтологическая семантика М.: Физматлит, 2013. 348 с.
 10. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. М.: Изд-во ИНТУИТ, 2008. 176 с.
 11. Боровикова О.И., Загоруйко Ю.А., Загоруйко Г.Б., Кононенко И.С., Соколова Е.Г. Разработка портала знаний по компьютерной лингвистике // Труды 11-й национальной конференции по искусственному интеллекту с международным участием КИИ-2008 (г. Дубна, Россия). Т. 3. М.: ЛЕНАНД, 2008. С. 380–388.
 12. Крылов С.А. Из каких элементов состоит метаязык лингвистики? // Кибрик А.Е. (гл. ред.). Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции “Диалог” (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 248–253.
 13. Антопольский А.Б., Ефременко Д.В. Инфосфера общественных наук России: монография / под ред. В. А. Цветковой. М.; Берлин: Директ-Медиа, 2017. 676 с. DOI: 10.23681/468227
 14. Антопольский А.Б., Савчук С.О., Тамеев А.А. О разработке онтологии поисковых терминов по лингвистике // Информационные ресурсы России. 2020. № 4. С. 2–7.
 15. Антопольский А.Б., Максимов Н.В., Тамеев А.А. Экспериментальная база данных источников для создания онтологии по лингвистике // Информационные ресурсы России. 2021. № 3. С. 24–30. DOI: 10.46920/0204-3653_2021_03181_24
 16. Каленов Н.Е., Белоозеров В.Н. Формирование терминологических словарей по лексике классификационных систем // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. 2015. № 3. С. 60–70.
 - Humanities in the Digital Age. Mozhaeva, G.V. (ed.)). Tomsk, Izd-vo Tom. un-ta Publ., 2016. 120 p. (In Russ.)
 2. Antopolsky, A.B. *Infosfera tsifrovoy humanitaristiki: opyt analiza* [Infosphere of Digital Humanities: The Experience of Analysis]. *Informacionnye resursy Rossii* [Information Resources of Russia]. 2022, No. 1, pp. 30–38. DOI: 10.52815/0204-3653_2022_01185_30 (In Russ.)
 3. Antopolsky, A.B. *Informacionnaya infrastruktura socialno-gumanitarnykh nauk v Evrosoyuze* [Information Infrastructure of Social Sciences and Humanities in the European Union]. *Nauka i nauchnaya informatsiya* [Scholarly Research and Information]. 2021, No. 4 (1), pp. 18–32. DOI: 10.24108/2658-3143-2021-4-1-18-32 (In Russ.)
 4. *Tsifrovye humanitarnye nauki. Hrestomatiya* [Digital Humanities. Anthology]. URL: <http://lib3.sfu-kras.ru/ft/LIB2/ELIB/b71/free/i-531505996.pdf> (In Russ.)
 5. *Gumanitarnye nauki v epohu tsifrovizatsii* [Humanities in the Era of Digitalization]. *Videotranslyatsiya Obshchego sobraniya Otdeleniya istoriko-filologicheskikh nauk RAN* [Video Broadcast of the General Meeting of the Department of Historical and Philological Sciences of the Russian Academy of Sciences]. URL: <http://hist-phil.ru/events/427/> (In Russ.)
 6. Antopolsky, A.B. *Infosfera tsifrovoy humanitaristiki: struktura i tendentsii razvitiya* [Infosphere of Digital Humanities: Structure and Development Trends]. *Elektronnoe informacionnoe prostranstvo dlya nauki, obrazovaniya, kultury: materialy IX Mezhdunarodnoy nauchno-prakticheskoy konferentsii (g. Orel, 16–17 dekabrya 2021 g.)*. Nauch. red. i sost. V.V. Matveev, D.N. Gribov, red. kol.: A.A. Aksyuhin, S.N. Manko [Electronic Information Space for Science, Education, Culture: Materials of the IX International Scientific and Practical Conference (Orel, December 16–17, 2021). Scientific Ed. and Comp. V.V. Matveev, D.N. Gribov, Ed. Col.: A.A. Aksyukhin, S.N. Manko]. Orel, Orlovskij gos. in-t kultury Publ., 2022, pp. 5–11. (In Russ.)
 7. Riley, Jenn. Seeing Standards: A Visualization of the Metadata Universe DOI: [org/10.5683/SP2/UOHPVH](https://doi.org/10.5683/SP2/UOHPVH). Scholars Portal Dataverse, V3. URL: gWl/jicj8wtJm4Grmph7TQ== [fileUNF] (available: 01.04.2022).
 8. Antopolsky, A.B. *Lingvisticheskie informacionnye resursy: monografiya* [Linguistic Information Resources: Monograph]. Moscow, INION RAN Publ., 2022. 466 p. (In Russ.)
 9. Rubashkin, V.Sh. *Ontologicheskaya semantika* [Ontological Semantics]. Moscow, Fizmatlit Publ., 2013. 348 p. (In Russ.)
 10. Dobrov, B.V., Ivanov, V.V., Lukashevich, N.V., Solovjev, V.D. *Ontologii i tezaurusy: modeli, instrumenty, prilozheniya* [Ontologies and Thesauruses: Models, Tools, Applications]. Moscow, Izd-vo INTUIT Publ., 2008. 176 p. (In Russ.)

REFERENCES

1. *Digital Humanities: humanitarnye nauki v tsifrovuyu epohu. Pod red. G.V. Mozhaevoy* [Digital Humanities:

11. Borovikova, O.I., Zagorulko, YU.A., Zagorulko, G.B., Kononenko, I.S., Sokolova, E.G. *Razrabotka portala znaniy po kompyuternoj lingvistike* [Development of a Knowledge Portal on Computational Linguistics]. *Trudy 11-j nacionalnoj konferencii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2008 (g. Dubna, Rossiya)* [Proceedings of the 11th National Conference on Artificial Intelligence with International Participation CII-2008 (Dubna, Russia)]. Vol. 3. Moscow, LENAND Publ., 2008, pp. 380–388. (In Russ.)
12. Krylov, S.A. *Iz kakih elementov состоit metayazyk lingvistiki?* [What Elements does the Metalanguage of Linguistics Consist of?]. Kibrik, A.E. (gl. red.). *Kompyuternaya lingvistika i intellektualnye tekhnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog" (Bekasovo, 26–30 maya 2010 g.). Vyp. 9 (16)* [Kibrik, A.E. (Editor-in-Chief). Computational Linguistics and Intellectual Technologies: Based on the Materials of the Annual International Conference "Dialog" (Bekasovo, May 26–30, 2010). Issue 9 (16)]. Moscow, Izd-vo RGGU Publ., 2010, pp. 248–253. (In Russ.)
13. Antopolsky, A.B., Efremenko, D.V. *Infosfera obshchestvennykh nauk Rossii: monografiya. Pod red. V.A. Tsvetkovoj* [Infosphere of Social Sciences of Russia: Monograph. Tsvetkova, V.A. (Ed.)]. Moscow, Berlin, Direkt-Media Publ., 2017. 676 p. DOI: 10.23681/468227 (In Russ.)
14. Antopolsky, A.B., Savchuk, S.O., Tameev, A.A. *O razrabotke ontologii poiskovykh terminov po lingvistike* [On the Development of an Ontology of Search Terms in Linguistics]. *Informacionnye resursy Rossii* [Information Resources of Russia]. 2020, No. 4, pp. 2–7. (In Russ.)
15. Antopolsky, A.B., Maksimov, N.V., Tameev, A.A. *Eksperimentalnaya baza dannykh istochnikov dlya sozdaniya ontologii po lingvistike* [Experimental Database of Sources for Creating an Ontology on Linguistics]. *Informacionnye resursy Rossii* [Information Resources of Russia]. 2021, No. 3, pp. 24–30. DOI: 10.46920/0204-3653_2021_03181_24 (In Russ.)
16. Kalenov N.E., Beloozerov V.N. *Formirovanie terminologicheskikh slovarей po leksike klassifikatsionnykh sistem* [Formation of terminological dictionaries on the vocabulary of classification systems]. *Nauchno-tekhnicheskaya informatsiya. Ser.1. Organizatsiya i metodika informatsionnoy raboty* [Scientific and Technical Information. Ser. 1. Organization and Methodology of Information Work]. 2015, No. 3, pp. 60–70. (In Russ.)

Дата поступления материала в редакцию: 4 апреля 2022 г.

Статья поступила после рецензирования и доработки: 25 января 2023 г.

Статья принята к публикации: 15 февраля 2023 г.

Дата публикации: 30 апреля 2023 г.

Received by Editor on April 4, 2022

Revised on January 25, 2023

Accepted on February 15, 2023

Date of publication: April 30, 2023