

Границы и содержание инфосферы цифровой гуманитаристики

А.Б. Антопольский

Институт научной информации по общественным наукам РАН

Ale5695yandex.ru

Аннотация

Обсуждаются методические вопросы создания справочно-информационной системы по цифровой гуманитаристике, которую совместно проектируют ИНИОН РАН и Сибирский федеральный университет. Основные вопросы – это тематические и институциональные границы инфосферы данной дисциплины, а также состав объектов, которые целесообразно отражать в системе. Обсуждается создание русскоязычной онтологии по цифровой гуманитаристике, возможности для ее разработки различных источников.

Ключевые слова: цифровая гуманитаристика, инфосфера, тематика, виды информационных объектов, институции, онтология.

Библиографическая ссылка: Антопольский А.Б. Границы и содержание инфосферы цифровой гуманитаристики // Информационное общество: образование, наука, культура и технологии будущего. Выпуск 6 (Труды XXV Международной объединенной научной конференции «Интернет и современное общество», IMS-2022, Санкт-Петербург, 23 – 24 июня 2022 г. Сборник научных статей). — СПб.: Университет ИТМО, 2022. С. 62-70. DOI: 10.17586/2587-8557-2022-6-62-70.

1. Введение

В настоящее время в ИНИОН РАН совместно с Сибирским Федеральным университетом ведутся исследования, имеющие целью создание справочно-информационной системы по цифровой гуманитаристике (далее - СИС ЦГ¹). В рамках этих исследований был проведен анализ основных объектов цифровой гуманитаристики, создаваемых в центрах цифровой гуманитаристики ведущих зарубежных стран, прежде всего Европы и Северной Америки. Основные результаты этого анализа были опубликованы в работе автора [1].

На основе этого анализа предполагалось определить границы и содержание инфосферы, для которой целесообразно создать упомянутую справочную систему. Однако очевидно, что определение границ инфосферы зависит, с одной стороны, от задач СИС ЦГ, с другой – от трактовки самого понятия ЦГ.

Основные задачи справочно-информационной системы по цифровой гуманитаристике можно сформулировать следующим образом.

СИС ЦГ должна использоваться:

- при разработке программ НИР;
- при подготовке и экспертизе заявок на гранты, проекты, темы НИР,

¹ В англоязычной литературе наиболее распространен термин Digital Humanities – ДН. Он иногда переводится как цифровые гуманитарные науки, мы предпочитаем термин цифровая гуманитаристика (ЦГ).

- для оценки результатов НИР, а также учреждений, подразделений и отдельных ученых;
- для координации работ по оцифровке
- для поддержки образовательной деятельности в области гуманитарных и общественных наук.

СИС ЦГ должна содержать информацию об объектах ЦГ, создаваемых в России, в целях повышения их видимости, цитируемости и повторного использования.

Таким образом, основным режимом использования СИС ЦГ должен стать поиск аналогов для информационных ресурсов, инструментов и других продуктов, создаваемых в организациях ЦГ, с целью избежать дублирования или его обнаружить.

Что касается определения дисциплины или области знания, которую принято называть ЦГ, то известно множество попыток ее определения. Они собраны, в частности, в хрестоматии, изданной в Сибирском федеральном университете [2].

Напомним, что понятие инфосферы автор ранее определял как совокупность ресурсов, сервисов и институций, участвующих в процессах научной коммуникации [3]. В цитированном выше исследовании, с учетом предложенного определения данного понятия, перечень объектов инфосферы ЦГ эмпирически формировался как совокупность институций, относящихся к ЦГ, выполняемых ими проектов и программ, электронных (цифровых) ресурсов, создаваемых этими институциями, а также сервисов и инструментов, используемых в этой деятельности.

Отношение институций к ЦГ оформляется в виде их участия в нескольких международных ассоциациях ЦГ. В Евросоюзе созданы специальные инфраструктурные консорциумы, координирующие и поддерживающие деятельность в этой сфере, например DARIAH [4].

Примером институционального оформления зарубежной практики ЦГ на национальном уровне может служить французская программа ЦГ под названием Humanum [5], включающая 200 научно-образовательных учреждений. В рамках этой программы поддерживается около 350 сайтов разнообразных ресурсов, проектов, сервисов и нормативов,

Применительно к России такой подход оказался затруднительным. Дело в том, что, в отличие от зарубежных стран, число российских институций, непосредственно относящихся себя к ЦГ и входящих в профессиональную ассоциацию DH-Russia [6], невелико. Большинство проектов, аналогичных зарубежным проектам ЦГ, выполняется научными или образовательными учреждениями, или институтами памяти самого различного профиля.

Далее мы рассмотрим подробнее возможные подходы к определению границ российской инфосферы ЦГ по нескольким аспектам.

2. Тематические границы цифровой гуманитаристики

Начнем с наиболее очевидного аспекта – тематического. Само понятие гуманитарных наук в тематическом отношении не является строгим. Хотя в большинстве практических классификаций наук, гуманитарные науки и выделяются, разумного определения гуманитарных наук не существует, а их состав и соотношение с социальными науками значительно варьируется.

Например, невозможно объяснить, почему психология или педагогика обычно относятся к социальной сфере, а история или лингвистика – к гуманитарной. Кроме того, множество социогуманитарных дисциплин вообще не привязан к этим понятиям, например, библиотековедение или издательское дело.

Тематический состав массива в 4,2 тыс. объектов зарубежной ЦГ, выделенных в цитированном выше исследовании, и приведенный на табл. 1, показал, тем не менее, что в зарубежной ЦГ доминируют вполне определенные тематические дисциплины.

Это историко-филологические, культурно-искусствоведческие дисциплины, информатика (в широком смысле), коммуникации, науковедение, а также ЦГ в целом.

Такой набор дисциплин трудно соотнести с обычной группировкой гуманитарных наук, однако традиционно понимаемые социальные науки представлены гораздо хуже.

Таблица 1. Тематический состав пилотного массива ЦГ

Рубрика ГРНТИ	Кол-во объектов	Рубрика ГРНТИ	Кол-во объектов
00 Общественные науки в целом	260	20 Информатика	940
02 Философия	17	21 Религия Атеизм	65
03 История и исторические науки	594	23 Комплексное изучение отдельных стран и регионов	38
04 Социология	19	28 Кибернетика	152
05 Демография	15	36 Геодезия. Картография	6
06 Экономика и экономические науки	1	50 Автоматика Вычислительная техника	96
10 Государство и право. Юридические науки	22	60. Полиграфия. Репрография. Фотокино-техника	2
11 Политика и политические науки	36	67 Строительство. Архитектура	14
12 Науковедение	95	68 Сельское и лесное хозяйство	2
13 Культура. Культурология	215	76 Медицина и здравоохранение	10
14 Народное образование. Педагогика	96	78 Военное дело	11
15 Психология	7	82 Организация и управление	2
16 Языкознание	877	83 Статистика	4
17 Литература. Литературоведение. Устное народное творчество	241	84 Стандартизация	24
18 Искусство. Искусствоведение	193	87 Охрана окружающей среды	5
19 Массовая коммуникация. Журналистика. Средства массовой информации	157		

На семинаре, на котором обсуждались результаты данного исследования [7], было принято решение, что на первом этапе создания СИС ЦГ основными будут историко-филологические и информационно-коммуникационные дисциплины. В дальнейшем тематические границы СИС ЦГ должны быть расширены.

3. Институциональный состав российской ЦГ

В зарубежной ЦГ основными акторами являются университеты или их подразделения, которые составляют подавляющую часть институций ЦГ. В России ситуация другая. Проекты и ресурсы, которые естественно отнести к сфере ЦГ, выполняются чаще всего библиотеками, архивами, музеями, академическими учреждениями.

Поэтому встает вопрос, как в СИС ЦГ должны быть представлены ресурсы и проекты институты памяти, которых в Интернете достаточно много. Промежуточное решение было принято следующее: в СИС ЦГ включаются ресурсы институтов памяти федеральных и субъектов РФ. Что касается образовательных структур социального и гуманитарного профиля то они и их ресурсы должны включаться в СИС ЦГ в случае их уникальности или оригинальности.

В дальнейшем список учитываемых институций и их продуктов и ресурсов может быть существенно расширен.

4. Типология объектов ЦГ

В пилотном исследовании, на которое мы ссылаемся, учитывались следующие типы объектов ЦГ:

- организации (институции) ДН, включая ассоциации, учреждения, консорциумы и исследовательские коллективы;
- информационные ресурсы, создаваемые участниками проектов и программ в сфере ДН;
- инструментальные (программные) средства, создаваемые в результате исследований или используемые в проектах в сфере ДН;
- сервисы, реализуемые для обслуживания сферы ДН;
- нормативно-технологические средства, используемые при создании ресурсов и сервисов ДН, включая стандарты, форматы, системы метаданных, языки разметки;
- прочие проекты в сфере ДН, которые не могут быть отнесены к вышеперечисленным типам информационных объектов.

Распределение этих объектов ЦГ, приведено на табл. 2.

Таблица 2. Распределение объектов ЦГ по типам

Тип объекта	Число объектов	Доля в %
Институции	696	17
Ресурсы	1667	40
Инструменты	1341	32
Проекты	201	5
Нормативы	142	4
Сервисы	72	2

Для эксперимента такая типология была достаточна. Однако для практического создания СИС ЦГ состав типов объектов требует определить более строго и решить многие методические вопросы. Например, относятся ли к институциям физические лица-авторы проектов? Или: как разделить инструменты и нормативы? Например, языки разметки, это что?

В любом случае, очевидно, что разделение объектов ЦГ на 6 типов является слишком грубым. По крайней мере, этой типологии недостаточно для поиска аналогов, что является, вероятно, основной задачей проектируемой информационной системы. Необходимо также определить более конкретный вид объекта. При этом исходное разделение на основные типы является весьма существенным, поскольку видовое деление существенно различается для разных типов.

Существует рекомендуемый перечень видов объектов ДН, разработанный в рамках таксономии ЦГ под названием TADIRAN группой европейских экспертов [8]. Этот перечень представлен на табл. 3.

Этот перечень применяется некоторыми каталогами в сфере ЦГ [9, 10]. Однако вопрос о его непосредственном применении для рассматриваемой информационной системы пока отложен. Это связано с неопределенностью и сложностью различения некоторых видов объектов данной таксономии. См. например Проекты, Исследования, Процесс исследования, Результаты исследований. Неясно соотношение понятий, например, Программное обеспечение и Инструменты.

Таблица 3. Типы объектов ДН в таксономии TADIRAH

Artifacts	Артефакты
BibliographicListings	Библиографические списки
Computers	Компьютеры
Curricula	Учебные планы
DigitalHumanities	Цифровая гуманитаристика
Data	Данные
File	Файл
Images	Изображения
Images(3D)	Изображения(3D)
Infrastructure	Инфраструктура
Interaction	Взаимодействие
Language	Язык
Link	Связь
Literature	Литература
Manuscript	Рукопись
Map	Карта
Metadata	Метаданные
Данное Methods	Методы
Multimedia	Мультимедиа
Multimodal	Мультимодальные объекты
NamedEntities	Именованные сущности
Persons	Лица
Projects	Проекты
Research	Исследования
ResearchProcess	Процесс исследования
ResearchResults	Результаты исследований
SheetMusic	Музыкальные ноты
Software	Программное обеспечение
Sound	Звуки
Standards	Стандарты
Text	Текст
TextBearingObjects	Объекты, содержащие текст
Tools	Инструменты
Video	Видео
Visualization	Визуализация
VREs	Виртуальная исследовательская среда

Также в этом перечне в едином ряду представлены понятия самых разных категорий, от тематических рубрик до абстрактных понятий, что кажется не очень удобным для представления данных.

5. Видовой состав объектов ЦГ

В описываемом эксперименте объекты ЦГ (кроме институций) были заиндексированы ключевыми словами (КС), которые в дальнейшем могут использоваться в качестве тегов как поисковый инструмент. Конечно, часть КС совпадает с перечнем типов TADIRAH, но получившийся перечень больше, он включает около 200 КС.

В табл. 4 приводится алфавитный перечень наиболее частотных КС ($N \geq 20$) с указанием типов информационных объектов и числа объектов, заиндексированных этими КС.

Конечно, вопрос о включении в СИС ЦГ объектов разного вида требует более глубокого изучения и разработки специального методического аппарата.

Рассмотрим, например, такой вид ресурсов как электронные библиотеки (ЭБ). К этому виду могут быть отнесены самые разнообразные собрания или коллекции произведений – от информационных систем с богатыми функциональными возможностями до случайных подборок цифровых копий документов. Часто к ним относят, например, коллекции собственных изданий учреждения. Недаром существующие каталоги российских ЭБ имеют значительный разброс – от нескольких десятков до 4 тыс.

В создаваемой пилотной версии ЦГ в отношении ЭБ был принят совмещенный критерий: во-первых, включались ЭБ, создаваемые учреждениями, входящими в определенную нами инфосферу ЦГ, во-вторых, имеющие посещаемость и видимость выше эмпирического порога.

Таблица 4. Основные виды объектов ДН

КС	Тип ИО	N	КС	Тип ИО	N
3D	ИР	21	Персональные коллекции	ИР	78
NLP	Инстр	85	Порталы	ИР	36
Анализ речи	Инстр	20	Программное обеспечение	Инстр	31
Анализ сетевых данных	Инстр	20	Региональные коллекции	ИР	20
Аннотирование	Инстр, Норм	74	Регистр ресурсов	ИР	36
Археологические коллекции	ИР	83	Репозиторий	ИР	40
Архивные коллекции	ИР	55	Сайты проектов	Проекты	31
Аудио	ИР	27	Сбор контента	Инстр	23
Библиографии	ИР	41	Связанные данные	ИР, Инстр	30
Визуализация	Инстр	78	Словари	ИР	46
Геоданные	ИР	51	Статистический анализ	Инстр	38
ГИС	Инстр	26	Таксономии, онтологии	ИР	45
Изображения	ИР	20	Тематические коллекции	ИР	86
Инфраструктура	Проекты	20	Управление данными	Инстр	54
ИПС	Инстр	80	Управление документами	Инстр	23
Конкордансы	ИР, Инстр	23	Управление ресурсами	Инстр	30
Контент-анализ	Инстр	41	Учебные ресурсы	ИР	43
Корпуса	ИР	48	Форматные конверторы	Инстр.	23
Машинный перевод	Инстр, Сервис	20	Цифровая музыка	ИР	21
Метаданные	Норм	25	Цифровые издания	ИР	87
Музыкальные коллекции	ИР	97	ЭБ	ИР	101
Памятники, рукописи	ИР	96	Языки разметки	ИР	41
Перечни лиц	ИР	29	Языковые БД	ИР	117
Периодика	ИР	38			

Это только один пример. Но подобные критерии предстоит разработать для множества видов ресурсов.

6. О разработке онтологии ЦГ

Полученный в ходе эксперимента словарь КС, описывающих объекты ЦГ может быть одним из источников для возможного создания онтологии ЦГ. Такая онтология должна стать общим инструментом не только для индексирования объектов ЦГ, но и для любых других российских проектов и ресурсов.

Другим обязательным источником, в целях согласования с зарубежными проектами, может стать таксономия TADIRAH, цитированная выше. Перевод этой таксономии на русский язык, сделанный автором, опубликован в [11]. Хотя к этой таксономии у российских исследователей могут быть терминологические и содержательные претензии, TADIRAH вероятно, содержит самое полное представление понятий ЦГ в части классификации назначений проектов ЦГ и методов их реализации.

Конечно, существенное значение имеет авторитет коллектива разработчиков TADIRAH, а также нормативный и консенсусный характер этой таксономии.

Ценным источником для разработки онтологии по ЦГ может стать проект по визуализации пространства метаданных по культурному наследию, реализованный Jenn Riley [12]. В этом проекте собраны, систематизированы и представлены в виде наборов данных св. 100 стандартов, тезаурусов, классификаций, онтологий и других средств представления знаний в области ЦГ.

Если говорить о технологических средствах создания онтологии ЦГ, то это очевидно должен быть OWL [13].

Но кроме стандарта на технологию, важно, конечно, иметь концептуальную схему создания такой онтологии. В качестве примера разработки и реализации онтологии можно привести опыт реализации проектов в области ЦГ, имеющих целью интеграцию данных и метаданных, посвященных определенному историческому событию с использованием технологии связанных открытых данных. Подробное описание действий по разработке концептуальной модели онтологии, ее лексикализации, нормирования, сериализации и представления в виде графа знаний можно найти, например, в проекте [14]. Обзор подобных проектов сделан автором в работе [15].

7. Заключение

В качестве главного вывода можно сделать следующий: если за рубежом дисциплина ЦГ приобрела конкретную форму и содержание и уже сформировалась специальная инфосфера ЦГ, то в России эта дисциплина находится в стадии становления. Ее границы не определены ни в тематическом, ни в институциональном аспектах. Важно, конечно, что в России отсутствует национальная программа ЦГ, а также инфраструктура этой деятельности.

Предлагаемый проект СИС ЦГ может стать одним из элементов столь необходимой для нашей страны инфраструктуры ЦГ. Очевидно, однако, что для его грамотной реализации требуется деятельное участие многих специалистов из различных учреждений, заинтересованных в выработке консенсуса по кардинальным вопросам становления и развития ЦГ России.

Литература

- [1] Антопольский А.Б. Инфосфера цифровой гуманитаристики: опыт анализа // Информационные ресурсы России. 2022. № 1. С. 30-38. DOI: 10.52815/0204-3653_2022_01185_30.
- [2] Цифровые гуманитарные науки Хрестоматия. URL: <http://lib3.sfu-kras.ru/ft/LIB2/ELIB/b71/free/i-531505996.pdf> (дата обращения: 22.05.2022).
- [3] 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.
- [4] Digital Research Infrastructure for the Arts and Humanities. URL: <https://de.dariah.eu/> (дата обращения: 22.02.2022).
- [5] Très Grande Infrastructure de Recherche des humanities numerique. URL: <https://www.huma-num.fr> (дата обращения: 22.02.2022).

- [6] Российская ассоциация цифровых гуманитарных наук. URL: <http://dhrussia.ru/> (дата обращения: 22.02.2022).
- [7] Антопольский А.Б. «Инфосфера цифровых гуманитарных наук». Доклад на семинаре «Цифровая среда» DHRI@SFU. Сибирского федерального университета 19.01.2022. URL: <https://www.youtube.com/watch?v=-LXs-LbieIg> (дата обращения: 22.02.2022).
- [8] Taxonomy of Digital Research Activities in the Humanities (TADIRAH). URL: <https://vocabs.dariah.eu/tadirah2/en/> (дата обращения: 22.02.2022).
- [9] DiRT (Digital Research Tools) directory. URL: <http://dirtdirectory.org/> (дата обращения: 22.02.2022).
- [10] DARIAH Teaching Resources Registry. URL: <https://de.dariah.eu/schulungsmaterialsammlung> (дата обращения: 22.02.2022).
- [11] Антопольский А.Б. Языки индексирования для цифровой гуманитаристики // Научно-техническая информация. Сер. 2. Информационные процессы и системы. 2022. № 1. С.1-9. DOI: 10.36535/0548-0027-2022-01-1.
- [12] Seeing Standards: A Visualization of the Metadata Universe. URL: <http://jennriley.com/metadatamap/> (дата обращения: 22.02.2022).
- [13] Web Ontology Language. URL: https://ru.wikipedia.org/wiki/Web_Ontology_Language (дата обращения: 22.02.2022).
- [14] The Stanford Prison Experiment on LOD. URL: <https://spelod.github.io/#erModel> (дата обращения: 22.02.2022).
- [15] Антопольский А.Б. Связанные открытые данные в цифровой гуманитаристике (обзор публикаций) // Научно-техническая информация. Сер. 2 Информационные процессы и системы. 2022. № 5. с. 23-30 DOI: 10.36535/0548-0019-2022-05-4.

The Boundaries and Content of the Digital Humanities Infosphere

A.B. Antopolskii

Institute of Scientific Information on Social Science of the Russian Academy of Science

Methodological issues of creating a reference and information system on digital humanities, which is jointly designed by INION RAS and Siberian Federal University, are discussed. The main issues are the thematic and institutional boundaries of the infosphere of this discipline, as well as the composition of objects that it is advisable to reflect in the system. The creation of a Russian-language ontology on digital humanities and the possibilities for its development from various sources are discussed.

Keywords: digital humanities, infosphere, topics, types of information objects, institutions, ontology.

Reference for citation: Antopolskii A.B. The Boundaries and Content of the Digital Humanities Infosphere // Information Society: Education, Science, Culture and Technology of Future. Vol. 6 (Proceedings of the XXV International Joint Scientific Conference «Internet and Modern Society», IMS-2022, St. Petersburg, June 23-24, 2022). - St. Petersburg: ITMO University, 2022. P. 62 – 70. DOI: 10.17586/2587-8557-2022-6-62-70.

Reference

- [1] Antopolskii A.B. Infosfera cifrovoj gumanitaristiki: opyt analiza // Informacionnye resursy Rossii. 2022. № 1. P. 30-38. DOI: 10.52815/0204-3653_2022_01185_30. (In Russian).

- [2] Cifrovye gumanitarnye nauki Hrestomatiya. URL: <http://lib3.sfu-kras.ru/ft/LIB2/ELIB/b71/free/i-531505996.pdf> (access date: 22.05.2014). (In Russian).
- [3] Antopolskii A.B., Efremenko D.V. Infosfera obshchestvennykh nauk Rossii: monografiya / pod red. V.A. Cvetkovoj. M.; Berlin: Direkt-Media, 2017. 676 p. DOI: 10.23681/468227. (In Russian).
- [4] Digital Research Infrastructure for the Arts and Humanities). URL: <https://de.dariah.eu/> (access date: 22.05.2014).
- [5] Très Grande Infrastructure de Recherche des humanities numerique. URL: <https://www.huma-num.fr> (access date: 22.05.2014).
- [6] Rossijskaya asociaciya cifrovyyh gumanitarnyyh nauk. URL: <http://dhrussia.ru/> (access date: 22.05.2014). (In Russian).
- [7] Antopolskii A.B. Infosfera cifrovyyh gumanitarnyyh nauk. Doklad na seminare «Cifrovaya sreda» DHRI@SFU. Sibirskogo federal'nogo universiteta 19.01.2022. URL: <https://www.youtube.com/watch?v=-LXs-LbieIq> (access date: 22.05.2014). (In Russian).
- [8] Taxonomy of Digital Research Activities in the Humanities (TADIRAH). URL: <https://vocabs.dariah.eu/tadirah2/en/> (access date: 22.05.2014).
- [9] DiRT (Digital Research Tools) directory. URL: <http://dirtdirectory.org/> (access date: 22.05.2014).
- [10] DARIAH Teaching Resources Registry. URL: <https://de.dariah.eu/schulungsmaterialsammlung> (access date: 22.05.2014).
- [11] Antopolskii A.B. Yazyki indeksirovaniya dlya cifrovoj gumanitaristiki Nauchno-tekhnicheskaya informaciya // Ser. 2 Informacionnye processy i sistemy. 2022 № 1. P. 1-9. DOI: 10.36535/0548-0027-2022-01-1. (In Russian).
- [12] Seeing Standards: A Visualization of the Metadata Universe. URL: <http://jennriley.com/metadatamap/> (access date: 22.05.2014).
- [13] Web Ontology Language. URL: https://ru.wikipedia.org/wiki/Web_Ontology_Language (access date: 22.05.2014).
- [14] The Stanford Prison Experiment on LOD. URL: <https://spelod.github.io/#erModel> (access date: 22.05.2014).
- [15] Antopol'skii A.B Svyazannye otkrytye dannye v cifrovoj gumanitaristike (obzor publikacij) // Nauchno-tekhnicheskaya informaciya ISSN 0548-0027. Ser. 2 Informacionnye processy i sistemy. 2022. № 5. p. 23-30 DOI: 10.36535/0548-0019-2022-05-4. (In Russian).