МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 1 Москва 2022

ИНФОРМАЦИОННЫЕ ЯЗЫКИ

УДК [7/9:004]:025.4.025/.026

А.Б. Антопольский

Языки индексирования для цифровой гуманитаристики

Обсуждается возможность использования в качестве средств индексирования и поиска для справочно-информационной системы по цифровой гуманитаристике (DH) специализированной таксономии по DH. Описывается применение ГРНТИ для описания тематической структуры DH и делается вывод о необходимости его существенной доработки для этой цели. Предлагается типология объектов DH. Приводится список ключевых слов, полученных при индексировании экспериментального массива объектов DH.

Ключевые слова: *Цифровая гуманитаристика, средства индексирования, TADIRAH, тематическая структура, ГРНТИ, типология, ключевые слова*

DOI: 10.36535/0548-0027-2022-01-1

ВВЕДЕНИЕ

В Институте научной информации по общественным наукам (ИНИОН РАН) начато исследование, конечная цель которого создание информационно-справочной системы по цифровой гуманитаристике. В ходе этой работы был создан экспериментальный массив объемом св. 4 тыс. объектов цифровой гуманитаристики. Методика и основные результаты этого исследования представлены в работе[1].

[1] А.Б. Антопольский. Инфосфера цифровой гуманитаристики: опыт анализа // Информационные ресурсы России. – 2022. – №1 – (в печати).

В настоящей статье рассматриваются вопросы выбора или создания языковых средств, которые могут использоваться для индексирования и поиска информационных объектов в этой системе.

Цифровая гуманитаристика (*Digital humanities* – DH) – это область научной деятельности на стыке компьютерных или цифровых технологий и гуманитарных дисциплин, включающая создание и систематическое использование цифровых ресурсов в гуманитарных науках, а также анализ их применения.

DH охватывает целый ряд тем от создания онлайн-коллекций первичных источников (текстовых, графических, аудиовизуальных) до интеллектуального анализа больших наборов культурных данных и тематического моделирования и включает как оцифрованные, так и рожденные в цифровом виде материалы, сочетая методологии из традиционных гуманитарных и отчасти социальных дисциплин, а также разнообразные методы и инструменты, предоставляемые современным уровнем развития информатики [1].

ТАКСОНОМИЯ TADIRAH

Очевидно, что цифровая гуманитаристика носит комплексный междисциплинарный характер, поэтому средства индексирования для этой сферы должны интегрировать понятия из разных областей. Европейские исследователи DH коллективными усилиями разработали специальную таксономию, получившую название TADIRAH, для представления информационных объектов инфосферы DH.

«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» [2].

TADIRAH включает 3 фасета:

- Виды исследовательской деятельности.
- Объекты исследования.
- Методы исследований.

Содержание этих фасетов (кроме толкования рубрик) на русском языке приводится в *Приложении 1*.

Фасет *Виды деятельности* имеет 2 уровня иерархии, остальные фасеты плоские. *Методы* частично привязаны к видам деятельности; методы, не привязанные, выделены отдельно. Перечень объектов образует отдельный фасет.

Анализ TADIRAH 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 описание этого вида деятельности по TADIRAH:

«*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 о*богащению* и *анализу*, по крайней мере, с систематической точки зрения, если не на практике».

Рубрика *Захват* включает такие виды деятельности, как *Преобразование, Распознавание данных, Раскрытие, Сбор, Отображение, Запись Транскрипция*, а в качестве метода указан только *Обход вебстраниц.*

Процитированное 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 такие виды деятельности как *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 TADIRAH отнесен не к *Захвату*, а к *Созданию.*

Различные вопросы возникают при анализе и других видов деятельности и методов, включенных в TADIRAH, в целом их классификация представляется не очень удобной для практического индексирования.

Более привычен для российского исследователя перечень типов объектов TADIRAH. Однако и здесь есть вопросы. 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 категорий, от тематических рубрик до абстрактных понятий, что кажется не очень удобным для индексирования.

Таким образом, вопрос о применении таксономии TADIRAH для проектируемой информационной системы по цифровой гуманитаристике пока отложен.

В то же время следует отметить, что TADIRAH практически уже применяется во многих каталогах и порталах DH. Перечень приложений TADIRAH можно найти, например, по адресу: Places where TaDiRAH is available, Initiatives using TaDiRAH. – URL: https://github.com/dhtaxonomy/TaDiRAH/blob/master/readme.md (дата обращения: 15.11.2021).

ТЕМАТИЧЕСКАЯ СТРУКТУРА СФЕРЫ ЦИФРОВОЙ ГУМАНИТАРИСТИКИ

Наиболее интересный и спорный вопрос при изучении инфосферы DH – это ее тематическая структура. Авторы многочисленных обзорных и аналитических публикаций по проблемам DH практически не обсуждают этот аспект. В TADIRAH тематический фасет отсутствует.

В то же время тематический фасет необходим для описания инфосферы цифровой гуманитаристики. Действительно, распространяется ли DH на все гуманитарные и/или социальные науки или только на некоторые? Относятся ли к сфере DH библиотечные, музейные, архивные издательские цифровые технологии, полностью или частично? Следует ли вклю-

чать в состав DH любые работы по компьютерной лингвистике, электронные библиотеки, аналитику социальных сетей и многие другие направления?

Мы считаем, что умозрительные ответы на эти вопросы дать невозможно, поэтому в предлагаемом нами исследовании рассматривается фактическая тематика проектов DH, как она складывается в действующих международных или национальных программах цифровой гуманитаристики.

В качестве примера в *Приложении 2* приводится перечень дисциплин [3], к которым отнесены ресурсы, созданные в рамках французской национальной программы DH под названием Huma-Num. Сама эта программа размещается по адресу: Très Grande Infrastructure de Recherche des humanities numerique. – URL: https://www.huma-num.fr (дата обращения 15.11.2021). Заметим, что эта программа – одна из самых больших: в ней участвует более 200 организаций, создано свыше 300 сайтов.

Легко видеть, что тематическая структура Huma-Num шире, чем, например, комплекс гуманитарных наук по применяемым в России классификациям, таким как ГРНТИ, FOS (известная также как классификация ОЭСР/OECD) [4] или номенклатура ВАК. В него входят не только гуманитарные, но и все социальные дисциплины, а также некоторые дисциплины, в российской традиции к социогуманитарным наукам не относящиеся (география, архитектура, статистика, исследования окружающей среды).

В нашем исследовании были обработаны объекты DH, входящие в международные и некоторые национальные программы и ассоциации или созданные в них. В таблице приводится использование ГРНТИ в качестве инструмента для тематического структурирования инфосферы цифровой гуманитаристики.

Заметим, что распределение результатов индексирования приводится только по одной рубрике ГРНТИ, хотя в нашем эксперименте часть объектов была отнесена к двум рубрикам.

В таблице видна крайняя неравномерность тематического распределения объектов DH.

Значительная доля (около 1,2 тыс.) объектов была отнесена к информатике в широком смысле, включая *20 Информатика*, *28 Кибернетика* (в части искусственного интеллекта) и *50 Автоматика и вычислительная техника* (в части программирования).

Распределение результатов индексирования объектов DH по первому уровню ГРНТИ

Код ГРНТИ	Рубрики	Количество объектов
00	Общественные науки в целом	260
02	Философия	17
03	История и исторические науки	594
04	Социология	19
05	Демография	15
06	Экономика и экономические науки	1
10	Государство и право. Юридические науки	22
11	Политика и политические науки	36
12	Науковедение	95
13	Культура. Культурология	215
14	Народное образование. Педагогика	96
15	Психология	7
16	Языкознание	877
17	Литература. Литературоведение. Устное народное творчество	241
18	Искусство. Искусствоведение	193
19	Массовая коммуникация. Журналистика. Средства массовой информации	157
20	Информатика	940
21	Религия Атеизм	65
23	Комплексное изучение отдельных стран и регионов	38
28	Кибернетика	152
28	Геодезия. Картография	6
50	Автоматика Вычислительная техника	96
60	Полиграфия. Репрография. Фотокинотехника	2
67	Строительство. Архитектура	14
68	Сельское и лесное хозяйство	2
76	Медицина и здравоохранение	10
78	Военное дело	11
82	Организация и управление	2
83	Статистика	4
84	Стандартизация	24
87	Охрана окружающей среды	5
	Всего	4120

Из собственно гуманитарных дисциплин основную долю составляют лингвистика, история, литература, искусство и культура. Всего к этим дисциплинам относятся 1,5 тыс. объектов. Заметим, что библиотечное, архивное и музейное дело относятся в ГРНТИ к рубрике *13 Культура. Культурология*.

Объекты, относящиеся к DH в целом (260 объектов, в основном институции), были отнесены нами к рубрике *00 Общественные науки в целом*, а электронные издания (150 объектов) – к рубрике *19 Массовая коммуникация. Журналистика. Средства массовой информации.*

Обращает на себя внимание небольшое количество объектов, связанных с социальными науками (экономика, психология, право, политология, социология, демография). Всего к социальным наукам было отнесено около 300 объектов, причем треть из них – это университеты, получившие код *14 Народное образование. Педагогика.*

Таким образом, хотя в большинстве программ цифровой гуманитаристики прокламируется применение DH как для гуманитарных, так и для социальных наук, на практике в программах DH доминирует историко-филологическая тематика, культура и искусство. В то же время в программах DH зачастую выполняются проекты, выходящие за пределы социогуманитарных наук, например, по медицине или экологии.

Заметим, что в таблице вообще отсутствует рубрика *26 Комплексные проблемы общественных наук.* Нынешнее содержание этой рубрики неактуально, хотя многие из важных понятий DH могли бы быть отражены именно в ней.

Как итог анализа полученного распределения можно констатировать, что в инфосфере DH преобладает тематика гуманитарной сферы и информационных технологий, хотя выход за пределы этой тематики не редкость.

ГРНТИ КАК ИНСТРУМЕНТ ОПИСАНИЯ ТЕМАТИКИ DH

Приведенные в настоящей статье данные относятся к распределению результатов индексирования по первому уровню ГРНТИ. Однако анализ на более глубоких уровнях показывает существенные недостатки ГРНТИ как инструмента для описания инфосферы цифровой гуманитаристики.

Прежде всего, это касается информатики. Во-первых, эта тематика, как мы уже отмечали, разнесена по трем рубрикам первого уровня, что, конечно, неудобно.

Во-вторых, и это главное, в ГРНТИ отсутствуют важнейшие понятия, определяющие современные информационные технологии, процессы, ресурсы и инструменты. Совершенно очевидно, что этот раздел ГРНТИ кардинально устарел. Можно напомнить, что он создавался в 70-х гг. XX века, когда понятие информационных технологий и их применения для различных методов обработки информации имело совершенно другое наполнение. Достаточно сказать, что в то время не было Интернета. Однако этот раздел ГРНТИ с тех пор подвергался лишь косметическим изменениям. Такой консерватизм объясним, поскольку изменение рубрикатора ведет в большинстве случаев к переиндексации фондов, что экономически и технологически нереализуемо. Однако пользоваться устаревшим рубрикатором для индексирования современных информационных ресурсов также невозможно.

В ходе нашего эксперимента было принято решение подготовить перечень существенных для инфосферы DH понятий, которые можно рассматривать в качестве потенциальных кандидатов для пополнения ГРНТИ в плане информационных технологий. Поскольку вопрос о включении этих понятий в иерархию ГРНТИ требует обсуждения, мы приводим список терминов в алфавитном порядке.

3D объекты
Визуализация
Генерация текстов
Извлечение из текста данных
Интеграция и компиляция контента
Инфографика
Контент-анализ
Метаданные
Обработка веб документов
Оцифровка
Платформы блогов
Ресурсы виртуальной и дополненной реальности
Связанные открытые данные
Семантическая сеть.
Сетевой анализ
Социальные сети
Сравнение информационных объектов
Средства коллективной работы
Текстовые редакторы
Технологии блокчейн
Управление документами,
Управление проектами
Форматы и форматные конверторы
Цифровые игры
Языки и инструменты разметки

Следует отметить, что в ходе эксперимента для индексирования использовалась модернизированная версия раздела ГРНТИ *16 Языкознание*, которая была разработана в 2019 г. специалистами ИНИОН РАН и ИРЯ РАН, согласована с ВИНИТИ РАН и официально представлена для внесения в ГРНТИ. Модернизированная версия представлена в монографии [5], а также размещена на сайте ИНИОН РАН [6]. Однако в эталонной версии ГРНТИ [7], размещенной на сайте ГПНТБ России, этот раздел представлен в старом неизмененном виде.

В целом новая версия этого раздела хорошо описывает объекты цифровой гуманитаристики. Тем не менее, опыт индексирования показал, что даже в новую версию раздела *16 Языкознание* имеет смысл ввести дополнения. Мы предлагаем в раздел *16.31. Прикладная лингвистика* ввести рубрики:

16.31.39 Языковые банки данных
16.31.43. Сентимент-анализ
16.31.45. Распознавание именованных сущностей.

Также считаем целесообразным дополнить и другие разделы ГРНТИ. Приведем общий список предлагаемых новых рубрик (кроме информатики и языкознания). Прежде всего, это введение понятия *цифровизации* в отдельных дисциплинах, для чего использовался типовой для этого понятия код xx.xx.85 (однако в рубрике *18.41. Музыка* этот код оказался занят):

00.85 Цифровая гуманитаристика
03.01.85 Историческая информатика
11.17 Политические репрессии. Холокост
13.51.85 Цифровые технологии в музейном деле
13.71.85 Цифровые архивы
13.73 Нематериальное культурное наследие
17.01.85 Цифровая филология
18.01.85 Цифровое искусство
18.31.85 Цифровые коллекции изобразительного искусства
18.41.47 Цифровые технологии в музыке
18.41.49 Цифровые музыкальные коллекции
18.45.85 Театр в цифровую эпоху
19.51.85 Цифровые издания
21.41.71 Религиозная и богослужебная литература
23.01.85 Цифровые ресурсы о географическом объекте.

Кроме введения новых рубрик, во многих случаях, можно обойтись изменениями формулировок названий, чтобы рубрика включала более современные или просто отсутствующие в ГРНТИ понятия. Приведем перечень предлагаемых изменений, в котором изменения выделены курсивом:

00.33 Терминология общественных наук. *Энциклопедии*
00.79 Кадры обществоведов. *Справочники персон*
04.51.67 Социология семьи и брака. Социология пола. *Гендерные проблемы*
13.61 Охрана памятников истории и культуры. *Культурное наследие*
16.21.36 Этнолингвистика. *Исчезающие языки*
16.31.33 Лингвистические вопросы искусственного интеллекта. *Онтологии*
19.21 Массовая коммуникация. *Реклама*
20.19.27 Автоматизация знаковой обработки текста. *Распознавание символов.*
20.51.23 Эффективность информационного обслуживания. *Оценка АИС*
20.53.17 Средства хранения информации. *Хостинг*
20.53.21 Средства выдачи информации. *Трансляция аудиовизуальных данных*
28.23.11 Языки представления и языки манипулирования знаниями. *Языки разметки*
28.23.24 Модели восприятия информации в интеллектуальных системах. *Извлечение данных*
50.41.23 Программное обеспечение вычислительных сетей. *Веб дизайн.*

Можно обратить внимание на общее соотношение ГРНТИ и инфосферы цифровой гуманитаристики. Всего для рубрицирования DH понадобилось около 400 рубрик из примерно 7 тыс., имеющихся в настоящее время в ГРНТИ. Например, из 400 рубрик раздела *10. Государство и право* использовано только 4 рубрики. Поэтому рубрикатор, более или менее равномерно отражающий тематическую структуру DH, должен иметь совсем иной состав.

Проведенное экспериментальное индексирование выявило и другие недостатки отражения в ГРНТИ социогуманитарной сферы. Например, архитектурная проблематика отделена от социогуманитарной сферы и вынесена в раздел *67 Строительство. Архитектура*. Военная история представлена в разделе *78. Военное дело* гораздо богаче, чем в профильном разделе *03 История*.

Заметим, что в описанном эксперименте большое количество объектов DH (около 200) достаточно условно отнесено к рубрике *03.29 История отдельных процессов, сторон и явлений человеческой деятельности*. Эта рубрика, конечно, требует детализации, что относится и к некоторым другим рубрикам. Для удобной навигации желательно иметь не более 20-30 объектов, относящихся к одной рубрике.

Типология объектов DH

Кроме тематической структуры для навигации и поиска объектов цифровой гуманитаристики представляется необходимым определить типологию этих объектов.

Нами были проанализированы каталоги и перечни различных объектов DH, созданных организациями, входящими в состав информационных сетей DH, таких как ADHO[2] или centerNet[3] и /или участников программ и инфраструктурных объединений, таких как Huma-Num или DARIAH[4]. В сферу исследования были включены также информационные объекты DH, прежде всего компьютерной лингвистики[5].

Проанализирован также перечень типов объектов DH, предлагаемых в таксономии TADIRAH.

В результате было принято решение на данном этапе для описания объектов цифровой гуманитаристики, кроме тематики, использовать 2 фасета: 1) основные типы объектов DH и 2) словари КС, представляющие виды объектов, их назначение и другие аспекты.

Перечень основных типов объектов DH получился следующим.

1. Институции DH, включая ассоциации, учреждения, консорциумы и исследовательские коллективы.
2. Информационные ресурсы, создаваемые в рамках программ DH.
3. Программные средства (инструменты), создаваемые в результате исследований или используемые в проектах DH.
4. Сервисы, реализуемые для обслуживания DH или поддерживающие важные для DH технологии.

[2] Alliance of Digital Humanities Organizations. – URL: https://wiki2.org/en/Alliance_of_Digital_Humanities_ Organizations (дата обращения 15.11.2021).
[3] An international network of digital humanities centers. – URL: https://dhcenternet.org/ (дата обращения 15.11.2021).
[4] The Digital Research Infrastructure for the Arts and Humanities (DARIAH). – URL: https://www.dariah.eu (дата обращения 15.11.2021).
[5] Антопольский А.Б. Лингвистические информационные ресурсы: монография. (Рукопись – одобрена к печати Ученым советом ИНИОН).

5. Нормативно-технологические средства, используемые при создании ресурсов и сервисов DH.

6. Прочие проекты в сфере DH, которые не могут быть отнесены к вышеперечисленным типам объектов.

На данном этапе каждый объект был отнесен к только одному типу. Этот подход определялся технологией отбора и дальнейшей организации БД. В некоторых случаях это приводило к дискуссионным решениям. В перспективе в спорных случаях объект можно будет относить к двум типам, тем самым снимая неопределенность за счет дублирования.

Конечно, разделение объектов DH на 6 типов является слишком грубым. По крайней мере, этой типологии недостаточно для поиска аналогов, что является, вероятно, основной задачей проектируемой информационной системы. Необходимо определить вид объекта более конкретно. При этом исходное разделение объектов DH на основные типы является важным и полезным, поскольку их видовое деление существенно различается для разных типов.

В нашем эксперименте тип *Институции* был разделен только по странам и выделены международные и европейские институции; к типу *Нормативы* были отнесены конкретные виды объектов: стандарты, методики, форматы, метаданные, языки разметки; типы *Программные инструменты* и *Сервисы* могут быть разделены на виды в соответствии с их функциональным назначением. Этот фасет структуризации инфосферы DH требует отдельного рассмотрения. При этом нужно иметь в виду, что он достаточно подробно разработан в таксономии TADIRAH, и имеется опыт его применения в больших каталогах, например, в каталоге инструментов DH TAPOR [8]. Однако предлагаемая в TADIRAH классификация видов деятельности (назначения) не может считаться общепринятой и использоваться напрямую. Интересно, что разработчики каталога TAPOR не ограничились классификацией TADIRAH и предложили еще и собственный перечень функций инструментов цифровой гуманитаристики.

Наиболее интересными и разнообразными типами объектов DH являются *информационные ресурсы* и *проекты* DH. Именно для их индексирования потребовались разнообразные понятия, обозначающие как виды информационных ресурсов, так и различные технологии, процессы, дисциплины, области применения.

Результат индексирования 3,5 тыс. объектов (кроме институций), представлен в виде словаря ключевых слов. На его основе может быть разработан тезаурус или облако тегов для организации поиска. Возможна также разработка таксономии DH (со словарями видов ресурсов и назначения инструментов и сервисов), согласованной с таксономией TADIRAH. Эта работа должна проводиться одновременно с доработкой ГРНТИ, если будет принято решение об использовании ГРНТИ так, чтобы понятия, обозначающие дисциплины, были включены в ГРНТИ и не дублировались в тезаурусе. Вместо ГРНТИ в проектируемой информационной системе возможна также разработка отдельной тематической классификации или тематического фасета таксономии DH.

В любом случае полученный массив ключевых слов дает большие возможности для анализа инфосферы цифровой гуманитаристики и разработки языковых средств структуризации этой сферы.

ЗАКЛЮЧЕНИЕ

Проведенное нами исследование позволило сделать следующие выводы:

- для создания информационно-справочной системы по цифровой гуманитаристике необходимо разрабатывать специализированное лингвистическое обеспечение, причем существующие языковые средства можно использовать лишь частично. В частности, тематический фасет является обязательным, однако применение для этой цели ГРНТИ требует существенной доработки и дополнения этого рубрикатора;
- понятийный аппарат DH нельзя признать общепринятым и устоявшимся, что показывает анализ таксономии TADIRAH;
- для организации навигации и поиска необходима типология объектов DH, задающая их видовой состав и функциональное назначение;
- полученный в ходе эксперимента словарь ключевых слов может быть использован для разработки специализированного тезауруса или организации поиска с помощью облака тегов.

СПИСОК ЛИТЕРАТУРЫ

1. Цифровые гуманитарные науки: хрестоматия / под ред. М. Террас, Д. Найхан, Э. Ванхутта, И. Кижнер: пер. с англ. – Красноярск: Сибирский. федеральный ун-т, 2017. – 352 с. – URL: http://lib3.sfu-kras.ru/ft/LIB2/ELIB/b71/free/i-531505996.pdf
2. TaDiRAH – Taxonomy of Digital Research Activities in the Humanities. – URL: http://tadirah.dariah.eu/vocab/index.php (дата обращения 15.11.2021).
3. Liste des sites web hébergés par Huma-Num. – URL: https://www.huma-num.fr/annuaire-des-sites-web/ (дата обращения 15.11.2021).
4. Расширенный классификатор OECD. – URL: https://www.vyatsu.ru/uploads/file/1703/kody_oecd_mezhdunarodnye.pdf (дата обращения 15.11.2021).
5. Антопольский А.Б. Научная информация и Электронное пространство знаний: монография / науч. ред. Д.Е. Ефременко. – Москва: ИНИОН РАН, 2020. – 252 с.
6. Рубрикатор информационных ресурсов по языкознанию. – URL: http://inion.ru/site/assets/files/1206/rubrikator.pdf (дата обращения 15.11.2021).
7. Государственный рубрикатор по научно-технической информации. – URL: https://www.gpntb.ru/images/2021/grnti/grnti2021.pdf (дата обращения 15.11.2021).
8. TAPoR (Text Analysis Portal for Research) collection. – URL: https://tapor.ca/home (дата обращения 15.11.2021).

Таксономия TADIRAH

Виды деятельности	**Методы**
1. Захват	
1.1. Преобразование	
1.2. Распознавание данных	
1.3. Раскрытие	
1.4. Сбор	
	Обход веб-страниц
1.5. Отображение	
1.6. Запись	
1.7. Транскрипция	
2. Создание	
2.1. Проектирование	
2.2. Программирование	
2.3. Перевод	
	Оцифровка
2.4. Веб-разработка	
2.5. Письмо	
3. Обогащение	
3.1. Аннотирование	
	Геопривязка
	Связанные открытые данные
	Распознавание именованных сущностей
3.2. Очистка	
3.3. Редактирование	
4. Анализ	
4.1. Контент-анализ	
	Информационный поиск
	Машинное обучение
	Распознавание именованных сущностей
	Сентимент-анализ
	Тематическое моделирование
4.2. Сетевой анализ	
4.3. Реляционный анализ	
	Распознавание образов
	Выравнивание последовательности
4.4. Пространственный анализ	
4.5. Структурный анализ	
	Анализ словосочетаний
	Конкорданция
	Машинное обучение
	Разметка частей речи
4.6. Стилистический анализ	
	Кластерный анализ
	Дистанционные измерения
	Машинное обучение
	Анализ основных компонентов
4.7. Визуализация	
5. Интерпретация	
5.1. Контекстуализация	
5.2. Моделирование	
5.3. Теоретизирование	
6. Хранение	
6.1. Архивирование	
6.2. Идентификация	
6.3. Организация	

Виды деятельности	Методы
6.4. Сохранение	Сохранение битового потока Долговечные постоянные носители Эмуляция Миграция Открытые архивные информационные системы Сохранение метаданных Репликация Сохранение технологий Управление версиями
7. Распространение	
7.1. Сотрудничество	
7.2. Комментирование	
7.3. Общение	
7.4. Краудсорсинг	Геймификация
7.5. Публикация	Связанные открытые данные
7.6. Разделение	
8. Мета-деятельность	
8.1. Оценка	
8.2. Создание сообщества	
8.3. Подготовка обзоров	
8.4. Управление проектом	
8.5. Преподавание и обучение	
	Методы, не связанные с видами Мозговая атака Просмотр Комментарии Отладка Кодирование Картирование Фотография Сканирование Поиск

Объекты DH

Артефакты	Метаданные
Библиографические списки	Методы
Взаимодействие	Музыкальные ноты
Видео	Мультимедиа
Визуализация	Мультимодальные объекты
Виртуальная исследовательская среда	Объекты, содержащие текст
Данные	Программное обеспечение
Звуки	Проекты
Изображения	Процесс исследования
Изображения(3D)	Результаты исследований
Именованные сущности	Рукопись
Инструменты	Связь
Инфраструктура	Стандарты
Исследования	Текст
Карта	Учебные планы
Компьютеры	Файл
Литература	Цифровая гуманитаристика
Лица	Язык

ПРИЛОЖЕНИЕ 2

Тематические дисциплины программы Huma-Num

Anthropologie sociale et ethnologie	Социальная антропология и этнология
Archéologie et Préhistoire	Археология и первобытное общество
Architecture, aménagement de l'espace	Архитектура и дизайн помещений
Art et histoire de l'art	Искусство и история искусства
Démographie	Демография
Droit	Право
Économies et finances	Экономика и финансы
Education	Образование
Etudes classiques	Классические исследования
Études de l'environnement	Исследования окружающей среды
Géographie	География
Héritage culturel et museology	Культурное наследие и музеология
Histoire	История
Histoire, Philosophie et Sociologie des sciences	История, философия и социология науки
Linguistique	Лингвистика
Littératures	Литература
Méthodes et statistiques	Методы и статистика
Musique, musicologie et arts de la scène	Музыка, музыковедение и театр
Psychologie	Психология
Religions	Религии
Science politique	Политология
Sciences de l'information et de la communication	Информация и коммуникация
Sociologie	Социология

Материал поступил в редакцию 17.11.21.

Сведения об авторе

АНТОПОЛЬСКИЙ Александр Борисович – доктор технических наук, профессор, главный научный сотрудник ИНИОН РАН, Москва
e-mail: ale5695@yandex.ru