ASEDU-2020

ITMO UNIVERSITY

Технологии и цифровизация: организация научных исследований в магистратуре

Центр исследований цифрового общества

Кононова Ольга Витальевна, канд.экон.наук

Университет ИТМО, СИ РАН, Санкт-Петербург

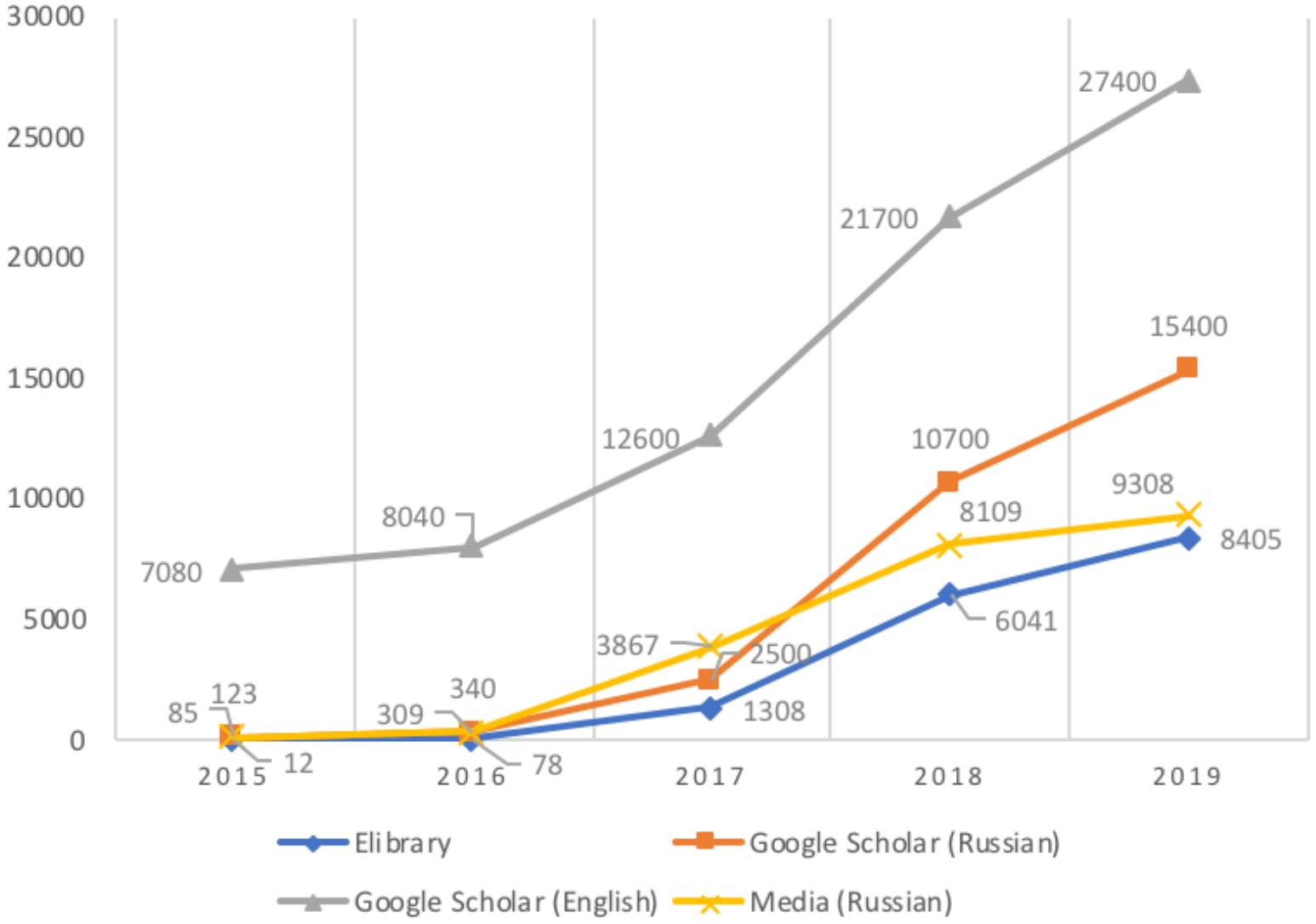
Прокудин Дмитрий Евгеньевич,

док.филос.наук, Санкт-Петербургский государственный университет, Университет ИТМО

БЛАГОТВОРИТЕЛЬНЫЙ ФОНД ВЛАДИМИРА ПОТАНИНА

Проект ГК200000654 2020

Динамика изменения потока научных публикаций и количества публикаций в СМИ. Запрос: «цифровая экономика» OR «digital economy»



БЛАГОТВОРИТЕЛЬНЫЙ ФОНД ВЛАДИМИРА ПОТАНИНА

УМК «Технологии извлечения и интеллектуального анализа данных в научных исследованиях»

направлен на формирование у магистрантов знаний и умений

- применения современных информационно-коммуникационных технологий (ИКТ) в научно-исследовательской и проектной деятельности
- комплексное использование технологий по поиску, извлечению и анализу научного знания. в открытых базах данных и научных источниках
- использование методик на базе технологий интеллектуальной обработки данных, систем продвинутого полнотекстового и мультимодального поиска, методов и инструментов извлечения контекстного знания с одновременным овладением широким спектром аналитического инструментария
- лучшей ориентации в развивающихся междисциплинарных научных областях, в которых терминологическая база еще не устоялась

УМК «Технологии извлечения и интеллектуального анализа данных в научных исследованиях» включает:

- учебно-методическое пособие и терминологический словарь;
- рекомендации по работе с информационной системой Научная электронная библиотека (http://elibrary.ru), другими научными информационными ресурсами и СМИ в интеграции с независимыми аналитическими системами;
- рекомендации по применению классов аналитических систем и аналитических сред T-Libra, Voyant-tools, Tropes High Performance Text Analysis, Sketch Engine, BigARTM, Mallet, систем машинного обучения и др.

УМК «Технологии извлечения и интеллектуального анализа данных в научных исследованиях»: ресурсы открытого доступа:

- репозиторий для размещения результатов выполняемых практических заданий в виде тематических коллекций контекстов и тезаурусов (на базе открытого программного обеспечения DSpace);
- агрегатор метаданных научных публикаций и иных информационных материалов (на базе открытого программного обеспечения Open Harvester Systems);
- электронный аннотированный каталог компьютерных систем для поддержки научных исследований и анализа контекстного знания, выделения и экспликации научного контента

Синтетический метод

- ✓ нацелен на извлечение контекстных знаний из неструктурированных или полуструктурированных информационных ресурсов
- ✓ позволяет посредством экспликации и картирования формировать коллекции релевантных тематике фрагментов (тематических контекстов)
- ✓ состоит в интеграции различных применяемых методов:
  - – отказ от изучения тематической выборки высокоцитируемых научных журналов с высоким импакт-фактором в пользу рассмотрения более широкого круга публикаций из тематически различных изданий
  - – синтез различных методик, интегральный охват инструментов исследования и варьирование последовательности применения технологий поиска, отбора, экспликации и анализа контекстного знания в зависимости от начальных условий и особенностей конкретного исследования
  - – реализация возможности создания и дальнейшего пополнения индивидуальных тематических коллекций тематических контекстов на основе кластеризации результатов запросов

Синтетический метод в УМК = формирование навыков комплексного использования современных ИКТ в решении задач поиска, извлечения, экспликации и анализа как научной, так и профессионально значимой информации. Через освоение метода и выполнение соответствующих учебных заданий формируются профессиональные компетенции исследователя и аналитика

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Синтетический метод: 3 фазы выявления и анализа контекстных знаний на практике

1 фаза

eLibrary: расширенный поисковый запрос, контекстный поиск, формирование корпуса текстов МДННИ

2 фаза

T-Libra: многослойный тематический запрос с вариацией используемых слоев, позволяющий выделять аспекты эксплицируемой темы; каскадный поиск (результаты одного запроса автоматически входят в поисковый образ другого запроса) для выявления новых элементов контекстного знания

3 фаза

eLibrary: полнотекстовый поиск; экспликация терминологического ландшафта МДННИ

Первая фаза - это поисковые запросы, терминологическое ядро и сборники текстовых материалов (тематические сборники), актуальные для междисциплинарного направления.

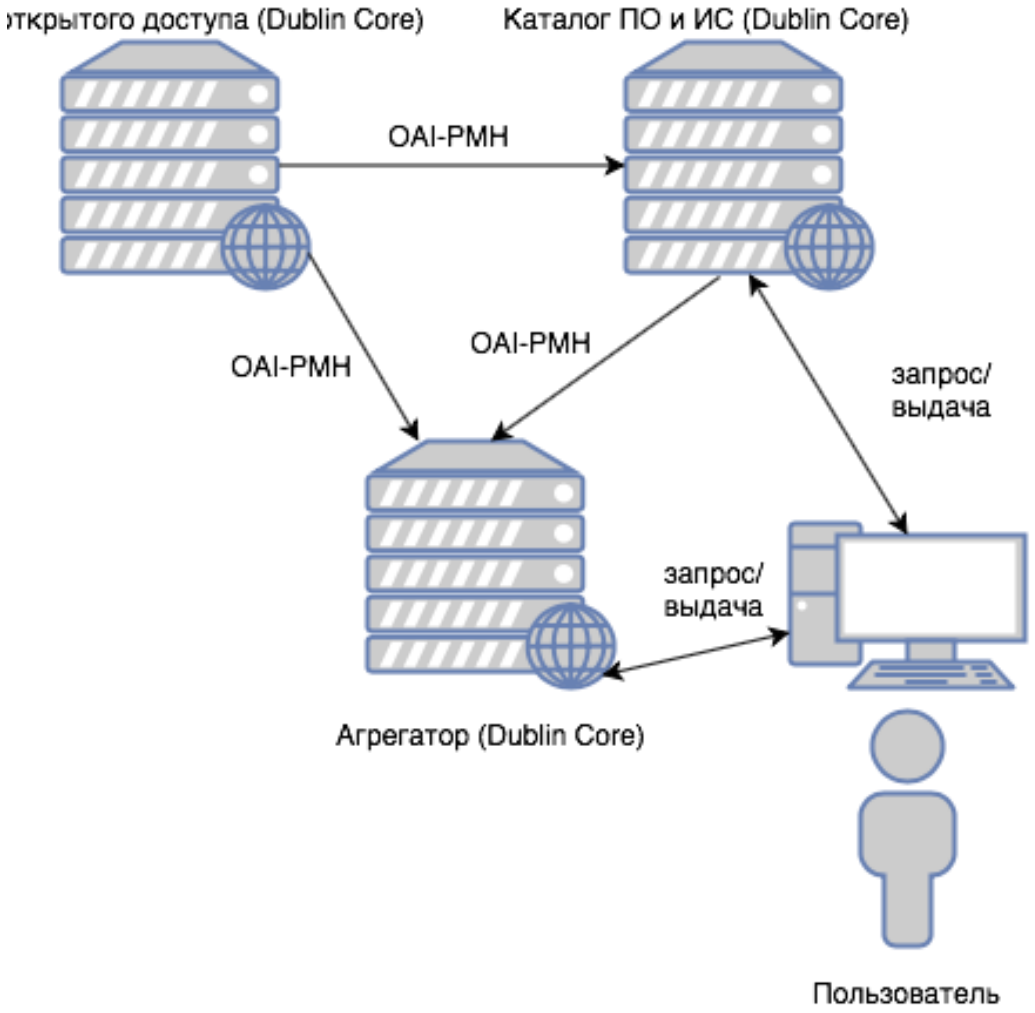
Вторая и третья фазы отражают последовательность шагов от выделения и экспликации первичных контекстов до построения предметно-тематических трендов в результате повторных тематических запросов в ранее эксплицированных массивах текстовых материалов.

УНИВЕРСИТЕТ ИТМО

Электронный аннотированный каталог компьютерных систем

Укрупнённые классы ПО и ИС в соответствии с основными видами обрабатываемых контекстов:

- Поисковые ИС **с** возможностью обработки большого объема неформализованных текстов (например, ИПС Яндекс, Google и т.п.)
- ИС, представляющие текстовые базы данных (например, eLibrary, T-Libra, Science Direct, Scopus, WoS и др.)
- Информационно-аналитические системы, обладающие возможностью обработки большого объема неформализованных текстов (например, Mallet, Voyant-Tools, Tropes, Sketch Engine, CLAVIRE и др.)
- Многофункциональные ИС смешанного типа, которые обладают достоинствами и недостатками ИС, описанных выше (например, ABBYY Intelligent Tagger SDK, ABBYY Smart Classifier SDK; Title: PROMT Analyser и др.).



Использование машиночитаемого каталога
в распределённой информационной среде

FIELD	VALUE
Title	Voyant-Tools
Creator	Sinclair, Stéfan Rockwell, Geoffrey
Subject	обработка отдельных документов; обработка коллекцией документов (корпус текстов); классификация; анализ интернет-страниц; частотный анализ; контекстный анализ; контекстуализация тенденций (построение трендов); визуализация данных анализа — термин; абзац; документ; коллекция документов
Description	Веб-ориентированная система для загрузки и анализа цифровых текстов, изучения частот и распределений терминов в документах и в коллекции документов (корпус). Представляет собой набор различных функциональных модулей. Существует локальное решение в виде приложения на JETTY.
Publisher	Информационные системы
Contributor	Andrew MacDonald Cyril Briquet Lisa Goddard Mark Turcato
Type	info:eu-repo/semantics/article info:eu-repo/semantics/publishedVersion — обработка текстов на естественном языке
Identifier	http://ojs.iculture.spb.ru/index.php/systems/article/view/5
Source	Информационные системы; Компьютерные программы и среды с функциями и сервисами извлечения и анализа контекстного знания для научных исследований
Language	rus
Relation	http://ojs.iculture.spb.ru/index.php/systems/article/view/5/2 http://ojs.iculture.spb.ru/index.php/systems/article/downloadSuppFile/5/2
Coverage	Web-ориентированное приложение (Web-интерфейс); Mac; Windows; JETTY server; Voyant server — —
Rights	(c) 2018 Voyant-Tools https://creativecommons.org/licenses/by/4.0/

Описание ПО и ИС схемой Dublin Core, полученное в агрегаторе OHS по протоколу OAI-PMH

Структурированное описание:

- использование спецификации метаданных Dublin Core
- основные характеристики описываются соответствующими элементами основного набора метаданных Dublin Core Metadata Element Set, DCMES
- описание видов контекстов – использование квалификаторов (расширение набора метаданных Dublin Core)

Предлагаемый набор метаданных:

dc.title — название ИС;
dc.creator — разработчик;
dc.subject.classification — основные функции (могут дополняться);
dc.subject.other — вид обрабатываемого контекста;
dc.description.abstract — описание ИС;
dc.publisher — издатель (правообладатель);
dc.contributor — внёсший вклад (люди/ организации: разработка ИС);
dc.date.issued — год последнего релиза;
dc.type — категории (классы ИС в соответствии с разработанной классификацией);
dc.format.mimetype — форматы обрабатываемых документов;
dc.identifier.uri — идентификатор (ссылка в сети Интернет на сайт разработчика);
dc.source.uri — источник (ссылка на веб-приложение);
dc.language — языки обрабатываемых документов;
dc.relation.isreferencedby — отношения (список публикаций по использованию ИС);
dc.coverage — поддерживаемые операционные системы;
dc.rights.license — тип лицензии.

Разработанный подход к представлению каталога ПО и ИС, предназначенных для анализа контекстного знания с функциями выделения, классификации и экспликации научного контента, на базе Dublin Core обеспечивает:

- интеграцию в каталог разработанной типологии контекстов, представляющей собой существенную характеристику, которая является основанием выбора ПО и ИС для проведения конкретных исследований;
- создание машиночитаемого каталога с использованием стандартного свободно распространяемого программного обеспечения (например, OJS, DSpace);
- эффективный поиск и отбор необходимых ПО и ИС для целей исследования в соответствии с основными характеристиками, описанными в тегах Dublin Core, используя стандартные поисковые механизмы;
- открытый доступ к элементам каталога как для пользователей, так и для автоматизированного индексирования;
- автоматизированный обмен по протоколу OAI-PMH для агрегации мета описаний каталога в других информационных системах.

Реализация проекта позволит
усовершенствовать учебный план программы магистратуры «Цифровые технологии умного города»
повысить аналитические и инструментальные компетенции преподавателей и сотрудников, выполняющих научное руководство магистрантами и аспирантами.

УМК будет использован в НИР магистрантов и аспирантов программ магистратуры Института дизайна и урбанистики ИТМО. Отдельные его элементы будут внедрены в образовательные программы подготовки бакалавров и магистров СПбГУ.
Университет и вузовское сообщество получат современный универсальный инструмент выполнения НИР, представленный до того фрагментарно в образовательной среде. Устойчивость и жизнеспособность результатов проекта состоит в универсальности подхода, независимости его от программного инструментария, что доказано авторами эмпирически и признано научным сообществом.

ITMO UNIVERSITY

СПАСИБО ЗА ВНИМАНИЕ!
THANK YOU FOR YOUR ATTENTION!

kononolg@yandex.ru

УНИВЕРСИТЕТ ИТМО