троицкий вариант наука
trv-science.ru

Запущен новый инструмент по поиску научной информации

19.11.2021 / Владимир Московкин / Один комментарий

Владимир Московкин

7 октября 2021 года был запущен новый инструмент по поиску научной информации The Global Index, разработанный архивистом Карлом Маламудом (Carl Malamud) под эгидой основанной им бесприбыльной корпорации Public Resource (Севастополь, Калифорния). Об этом широкой научной общественности в интервью с Карлом Маламудом рассказал 28 октября на австралийском портале Science alert ее журналист Дэвид Нилд (David Nield), а до этого 26 октября интервью с ним провела журналистка Холи Елсе (Holly Else) из Nature.

Этот индекс содержит около 107,2 млн журнальных статей, в общей сложности 38 терабайт данных в несжатом виде. Он охватывает более 355 млрд строк текста, каждая из которых содержит ключевое слово или фразу, взятую из опубликованной статьи. Практически это всё защищенные статьи коммерческих издателей. Это даже чуть больше, чем находится сейчас в базе данных Sci-Hub, насчитывающей около 88,5 млн статей.

Итак, мы имеем уже три ключевых инструмента по поиску научной информации в Интернете в виде научных статей: Google Scholar (запущен в 2004 году), Sci-Hub (запущен в 2011 году) и The Global Index (запущен в 2021-м). В чем их различия? Первый инструмент при поиске по ключевым словам охватывает практически полностью статьи из баз данных Scopus и Web of Science, но открывает полные тексты статей из журналов коммерческих издателей в случае, если они выложены в открытый доступ в виде авторских препринтов или файлов или в обход авторского права. Sci-Hub нелегально открывает практически любую статью из журналов коммерческих издателей при поиске по названию статьи или по ее DOI. Тогда встает вопрос, а зачем нужен еще один поисковый инструмент?

На это журналисту портала Science alert Дэвиду Нилду разработчик The Global Index Карл Маламуд ответил следующим образом: “Это поисковый инструмент, словарь знаний, карта знаний. Хотя мы уже упоминали Google, но он не совсем поисковая система. Ученым, использующим General Index (он же Global Index. — *Прим. автора*), придется кодировать свои собственные поисковые запросы для работы с ним. Скорее это тщательно каталогизированный и структурированный каталог, который можно использовать для изучения результатов научных исследований за десятилетия. Его основная цель — помочь с интеллектуальным анализом текста, что означает использование компьютеров для быстрого сканирования миллионов точек данных с целью поиска и перекрестных ссылок на что-то конкретное. Люди не могут прочитать и выделить ключевые данные из миллионов журнальных статей, но компьютерная программа, подключенная к General Index, может”.

В интервью журналу *Nature* в большей степени были затронуты юридические вопросы правомерности запуска этого инструмента. По мнению Маламуда, проект вполне укладывается в рамки закона: “Я очень уверен, что то, что я делаю, является законным”, — сказал Маламуд журналу *Nature*. Он также отметил, что делается это не для того, чтобы спровоцировать судебный процесс, а для развития науки. С другой стороны, некоторые юристы сомневаются в правомерности запуска этого инструмента, так как автор не разглашает сведения о том, получал ли он от издателей разрешение на скачивание их журнальных статей. Сам Карл Маламуд держит в тайне процедуру скачивания статей, убеждая всех, что в открытый доступ он не будет выставлять полные тексты статей, а только выдержки из них в окрестности ключевых слов. Это так называемые snippet, которые мы видим при работе с поисковой системой Google Books.

Редакция отдела новостей журнала *Nature* связалась с шестью издателями по поводу данного индекса. Все, кроме одного, отказались от комментариев. В заявлении Springer Nature говорится, что компания поддерживает инициативы в области открытой науки, в которых используются технологии и алгоритмы для удовлетворения потребностей исследователей. “Однако мы видели, как некоторые инициативы сталкивались с проблемами, когда не были обеспечены необходимые права для обеспечения их устойчивости”, — говорится в заявлении. Как видим, сделано очень обтекаемое заявление, авторы которого не в восторге от запуска нового индекса, в отличие от откликов самих ученых (этот издатель издает журнал *Nature*; отдел новостей *Nature* редакционно не зависит от своего издателя).

Сам The Global Index размещен на сайте Internet Archive по адресу archive.org/details/GeneralIndex, которому в этом году исполнилось 25 лет. На нем описана технология составления этого индекса, сконструированного из трех таблиц, полученных из 107 233 728 журнальных статей.

Первая таблица состоит из n-grams (в данном случае слов), ранжированных от unigrams (n = 1) до 5 – grams, выделяемых с помощью SpaCy. Каждая из 355 279 820 087 строк текста n-grams таблицы состоит из вышеуказанных n-grams (n изменяется от 1 до 5), соединенных с помощью DOI с журнальными статьями. Вторая таблица генерируется с помощью Yake и состоит из 19 740 906 304 строк, каждая из которых имеет ключевые слова и DOI статьи. Третья таблица связывает DOI статей с их метаданными.

Во избежание переполнения серверов весь корпус статей разбивается дважды на 16 кусков, в первом разбиении размещаются ключевые слова, во втором – n-grams.

После этого структурного описания данных приводится декларация в поддержку этого индекса, подписанная профессорами, независимыми исследователями, юристами, библиотекарями, менеджерами и компьютерными специалистами (100 подписей). На этой странице, созданной 7 октября 2021 года, есть ссылка на поисковый инструмент этого индекса. Простой поиск позволяет вести его по метаданным, ключевым словам, заголовкам телевизионных и радионовостей, а также по архивированным веб-сайтам.

Расширенный поиск достаточно сложен, он может проводиться с использованием специальных одиннадцати поисковых строк, а также с помощью возвращения JSON, XML, HTML, CSV и RSS форматов файлов. Ниже на веб-странице даны инструкции по проведению расширенного поиска.

В то же время в исследовательских и поисковых целях на первых порах вполне достаточен простой поиск. Проводя различные эксперименты в нем, мы выявили, что поиск идет в большей степени по книжным изданиям, чем по статьям. Причем выложены в открытый доступ старые издания, на которые по американскому законодательству истек срок авторского права (75 лет). Здесь кладезь сведений для историков, экономистов и других специалистов. Например, наберем в простом поиске (тип документа – text content) слово «НЭП», увидим 4837 результатов поиска, в левой полосе они распределены по типам файлов (больше всего текстовых файлов – 4808, фотографии часто соответствуют обложкам книг и журналов, которые раскрываются, и можно просматривать их тексты), годам (1923–2007, с ошибочным 1890 годом), темам (например, Kiev, Soviet history, Soviet Union), коллекциям (библиотечные и другие), создателям (издательства, организации, люди) и языкам (иностранные языки дают ссылки на словари).

Справа, в фиксированном состоянии, мы видим 8 обложек книг, документов, иногда журналов («Большевик», «Под знаменем марксизма», «Новая Россия», «Смена» и др.), под ними небольшие выдержки из текста или метаданных с выделением ключевого слова «НЭП». Все результаты поиска прокручиваются колесиком мышки, и вы в большем объеме видите всё разнообразие текстов.

В верхних результатах поиска мы обратили внимание на книгу «НЭП и кризис партии после смерти Ленина. Годы работы в ВСХН во время НЭП» (М.: Современник, 1991. — 367 с.). Это оказались мемуары Николая Владиславовича Валентинова (Вольского) (1880-1964), написанные в 1956 году, в которых анализ событий доведен до 1929 года. В русскоязычной «Википедии» отмечено, что эта книга является главным источником сведений для западных историков советской экономики, при этом ее англоязычное издание в англоязычной статье «Википедии» датируется 1971 годом.

Если кликнуть на обложку этой книги мышкой, то мы слева на темном фоне получим количество результатов поиска – 65, это говорит о том, что термин «НЭП» встречался в этой книги 65 раз, ниже приведены фрагменты текста с указанием страниц, в которых наблюдается этот термин. Справа на черном фоне дан разворот книги, который можно листать в обе стороны, увеличивая размер текста. Данную книгу можно прочесть целиком.

Ниже самой электронной книги приведены ее метаданные с аннотацией и содержанием книги, при этом год издания, язык, коллекция и тематика являются гиперссылочными, чтобы по ним происходил поиск и их сортировка. Под метаданными размещены различные идентификаторы, из которых мы видим, что книга сканирована интернет-архивом и помещена в рассматриваемый глобальный индекс 31.03.2019 г. Ниже имеется возможность оставить отзыв о книге (Reviews), а еще

ниже приведены гиперссылочные обложки книг по близкой тематике, выделяемые по метаданным. Всё то же самое мы увидим по любому другому изданию. В целом, очень удобный поисковый интерфейс, он намного лучше, чем в Google Books.

Продолжая просмотр изданий по данной тематике, мы протестировали термины “концессии” (2205 откликов, 1921–2020), “биржи” (9459 откликов, 1866–1998), “внешняя торговля” (1767 откликов, 1886–1988), “иностранный капитал” (710 откликов, 1922–2020). Все эксперименты с этим индексом были проведены 11.11.2021 г. По нашим наблюдениям, количество результатов поиска (откликов) в течение суток может меняться от 10 до 50 единиц.

Ниже по этим терминам мы привели по одной, наиболее любопытной для нас книге, в названии которой встречаются эти термины.

1. Дергачёва Н.П. Концессии. – Ленинград: Прибой, 1925. – 84 с. В работе концессии рассматриваются с точки зрения борьбы с капитализмом, приведены Декрет о концессиях, мнение Ленина о них при НЭПе, сравнение зарубежного и отечественного концессионного опыта, договор с английским капиталистом Лесли Уркартом (1874–1933). Много количественных данных по условиям концессионных договоров, предоставления сырьевой базы иностранной стороне и получения взамен готовой продукции и оборудования. Книга просматривается полностью, она отсканирована интернет-архивом и размещена в глобальном индексе 15.03.2019 г. Полный текст другого экземпляра этой книги выложен также на сайте Электронной библиотеки БЕЛИНКИ (библиотека Белинского) со штампом книгохранилища Свердловской областной библиотеки. Об авторе этой книги сведений в Интернете нет.
2. Биржи и рынки. Сборник. – М.: Бюро съездов биржевой торговли СССР, 1924. – 930 с. Том 1 открывается статьей Лазаря Моисеевича Когановича (1893–1991) “Ближе к биржам”, всего в этом сборнике представлено около 20 статей

российских экономистов, юристов и технических специалистов, многие из которых были репрессированы. Перечислены около сотни российских бирж, функционирующих во время НЭПа. Все страницы хорошо читаются, сборник отсканирован Интернет архивом и размещён в глобальном индексе 15.04.2019 г.

3. Внешняя торговля России в 1922–1923 хозяйственном году / под ред. В.Г. Громана и Л.Б. Кафенгауза. – М.: Экономическая жизнь, 1923. – 238 с. В сборнике представлены 11 статей советских экономистов и таблицы по внешней торговли. Первыми шли две статьи проф. Семёна Анисимовича Фалкнера (1890–1938) “Мировая конъюнктура к середине 1923 г.” и Михаила Яковлевича Кауфмана (1881–1946) “Обзор внешней торговли за первое полугодие 1922–1923 гг.” Потом шла статья Бориса Ефимовича Штейна (1892–1961) “Торговые договоры РСФСР”. Редакторы сборника Владимир Густавович Громан (1874–1940) и Лев Борисович Кафенгауз (1885–1940) также являлись известными в довоенное время экономистами. Все страницы хорошо читаются, сборник отсканирован интернет-архивом и размещен в глобальном индексе 09.03.2019 г.
4. Ляндау Л.Г. Иностранный капитал в дореволюционной России и в СССР. – М., Л.: Госиздат, 1925. – 84 с. Рассмотрены успехи НЭПа и концессионной политики, перечислены все 84 концессии, наибольшее их число зафиксировано для Германии – 22, Англия – 15, США – 10. Вся брошюра хорошо просматривается, она отсканирована интернет-архивом и размещена в глобальном индексе 18.10.2021 г. Автор — поляк Леон Германович Ляндау, родился в Лодзи в 1872 году, инженер-технолог, дата смерти неизвестна, в 1939 году приговорен к 15 годам ИТЛ, 5 лет поражения в правах (Ист. “Польские заключенные воркутинских лагерей”). Дополнительный наш поиск в Интернете показал, что он был заместителем председателя Концессионного комитета ВСХН РСФСР и первым директором НИИ резиновой промышленности (1930–1941, вторая дата, похоже, ошибочная). В «Википедии» приводятся все директора этого института с полными именами, за исключением Л.Г. Ляндау, его личное дело храниться в Российском государственном архиве экономики. Все эти данные не лежали на поверхности, и это хороший пример, как The Global Index может дать толчок к проведению историко-архивных изысканий.

Из других редких книг 1920–1940-х годов отметим еще три книги.

1. Терне А.М. В царстве Ленина: очерки современной жизни в РСФСР. – Берлин: Изд-во А.Терне, 1922. – 413 с. Эта книга об одном из первых свидетельств о жизни в Советской России. Экономист, статистик и социолог Андрей Михайлович Терне (1859–1921) в конце 1910-х годов был деканом экономического факультета Кубанского политехнического института, позже он провел несколько лет в Новороссийске, работая на советской службе и ведя жизнь рядового советского человека. Книга полностью просматривается, она отсканирована интернет-архивом и размещена в глобальном индексе 21.08.2021 г. В антикварном магазине “Русский библиофил” ее цена составляет 125 тыс. руб.
2. Knickerbocker H.R. Fighting the Red Trade Menace. – New York: Dodd, Mead and Company, 1931. — 295 p. Книга «Борьба с красной торговой угрозой» состоит из большого Введения и 24 глав по названиям европейских столичных и портовых городов, через которые шла торговля с Советским Союзом. Американский журналист голландского происхождения Хьюберт Ренфро Никербокер (1898–1949) за цикл статей о пятилетке в СССР, написанных для *Public Ledger,* в 1931 году удостоился Пулитцеровской премии. Книга просматривается полностью, она отсканирована интернет-архивом и

размещена в глобальном индексе 27.04.2019 г. Она продается в зарубежных книжных интернет-магазинах по цене около 1 тыс. долл.

3. За железной завесой: сборник материалов из мировой прессы / под ред. Г. Кремлёва. – Мюнхен: б.и., 1946. – 75 с. Проделан мониторинг советской и зарубежной прессы за 1946 год по более чем 60 статьям и репортажам. Все страницы хорошо читаются, сборник отсканирован интернет-архивом и размещен в глобальном индексе 31.01.2020 г. О редакторе этого сборника в Интернет ничего найти не удалось, хотя Google Books дает много ссылок на его репортажи и статьи, опубликованные в 1940–1950-е годы, особенно, по киноискусству.

В заключение следует сказать, что ученые, которые активно печатались в советское время за рубежом и в престижных советских научных журналах, могут увидеть свои имена через The Global Index в различных библиографиях, индексах, энциклопедиях и докладах, не подозревая об этом. Большой интерес представляют JPRS Report (1957–1995). Joint Publication Research Service действовала в период холодной войны как подразделение ЦРУ. Штат этой службы готовил переводы статей из разных стран и регионов мира, в том числе и из СССР (опубликовано за всё время более чем 130 000 докладов). Отметим, что в рассматриваемом индексе термин «холодная война» встречался 1132 раза, а “cold war” – 498 920 раз (12.11.2021 г.)

Владимир Московкин