ИПМ им.М.В.Келдыша РАН

Абрау-2021 • Труды конференции

З.В. Апанович

Сопоставление и интеграция информации о российских научных организациях из разноязычных источников данных

Рекомендуемая форма библиографической ссылки

Апанович З.В. Сопоставление и интеграция информации о российских научных организациях из разноязычных источников данных // Научный сервис в сети Интернет: труды XXIII Всероссийской научной конференции (20-23 сентября 2021 г., онлайн). — М.: ИПМ им. М.В.Келдыша, 2021. — С. 34-42.
https://doi.org/10.20948/abrau-2021-13
https://keldysh.ru/abrau/2021/theses/13.pdf

Видеозапись выступления

Размещена также презентация к докладу

Сопоставление и интеграция информации о российских научных организациях из разноязычных источников данных

З.В. Апанович[1]

[1]*Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук*

Аннотация. Информация о научных организациях является важным атрибутом, позволяющим идентифицировать авторов научных публикаций, а также анализировать географическое распределение публикаций и оценивать влияние на цитируемость публикаций, связанную с географическим фактором. К сожалению, информация о национальных научных организациях часто является неполной или искаженной в международных базах данных. Это касается, в частности, российских научных организаций, представленных в англоязычных базах данных. В данной работе представлены эксперименты по сопоставлению и интеграции информации о российских научных организациях в международных и российских источниках данных. Рассматриваются такие источники данных как GRID, Wikipedia, Wikidata и eLIBRARY.ru.

Ключевые слова: граф знаний, идентификация сущностей, корректность

Matching and integration of data about Russian research organizations from multilingual data sources

Z.V. Apanovich[1]

1 A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences

Abstract. Information about research organizations is an important attribute that enables identifying authors of scientific publications, as well as analyzing the geographical distribution of publications and assessing the impact on the citation of publications associated with a geographic factor. Unfortunately, information on national research-related organizations is often incomplete or

distorted in international databases. This applies, in particular, to Russian research organizations represented in English-language databases. The paper presents experiments on data matching and integration about Russian research organizations in multilingual data sources. Data sources such as GRID, Wikipedia, Wikidata and eLIBRARY.ru are considered.

Keywords: knowledge graph, identity resolution, correctness

Введение

Информация о научных организациях является важным атрибутом, позволяющим идентифицировать авторов научных публикаций [1, 2], а также анализировать географическое распределение публикаций и оценивать влияние на цитируемость публикаций, связанную с географическим фактором [3]. К сожалению, информация о национальных научных организациях, например, информация о российских научных организациях часто является неполной или искаженной в международных базах данных.

Одной из больших международных открытых баз данных научных организаций является база данных GRID [4] (Global Research Identifier Database, https://www.grid.ac/). Эта база данных содержит информацию о более чем ста тысячах научных организациях по всему миру и используется не только как самостоятельный источник данных. Данные GRID интегрированы в граф знаний SN SciGraph, разрабатываемый издательством Springer (https://www.springernature.com/gp/researchers/scigraph). Благодаря использованию данных GRID, все информация о публикациях SN SciGraph имеет географическую привязку (координаты, город, район, область, страна). Вторым достоинством GRID является наличие ссылок на глобальные библиографические ресурсы, такие как ROR (Research Organization Registry, htttps://ror.org), Crossref https://www.crossref.org/, и ISNI (International Standard Name Identifier, https://isni.oclc.org/). В настоящее время GRID содержит данные о 2019 исследовательских российских организациях. К сожалению, информация о российских научных организациях, хранящаяся в GRID, с одной стороны неполна, а с другой стороны, содержит явные неточности. Например, в GRID имеется страница, посвященная Сибирскому отделению Российской Академии Наук (Siberian Branch of the Russian Academy of Sciences, https://www.grid.ac/institutes/grid.415877.8). В качестве русскоязычного названия этой организации указан «Институт космофизических исследований и аэрономии им. Ю.Г. Шафера Сибирского отделения Российской академии наук», а в качестве дочерних организаций (“child institutes”) наряду с несколькими институтами, формально относящимися к СО РАН, перечислены образовательные организации различного подчинения, не имеющие к СО РАН никакого отношения. Среди них, например, имеется Восточно-сибирский институт Министерства

Внутренних дел Российской Федерации (East-Siberian Institute of the Ministry of Internal Affairs of the Russian Federation, (https://www.grid.ac/institutes/grid.445063.0), Сибирский юридический институт ФСКН России (https://www.grid.ac/institutes/grid.445537.4) и др. Также бросается в глаза значительное расхождение в количестве российских организаций GRID и количестве организаций самой большой базы данных российских научных организаций eLIBRARY.ru [5], на сайте которой представлено 12231 российских организаций. Так, если eLIBRARY.ru выдает список из 263 организаций, расположенных в Новосибирске, то GRID показывает только 75 таких организаций, причем некоторые из представленных организаций относятся к разряду прекративших существование.

На основе подобных примеров можно предположить, что русскоязычные источники данных должны содержать больше информации о российских организациях и информация должна быть лучшего качества чем англоязычные источники данных. Но прежде, чем сравнивать русскоязычные и международные источники данных, интересно посмотреть, как соотносятся данные о научных российских организациях в международных ресурсах.

Следует заметить, что в мире имеется значительное количество платформ, присваивающих идентификаторы научным организациям. ROR работает над присваиванием уникальных идентификаторов исследовательским организациям с 2018 года. Идентификаторы ROR использует в своих метаданных Crossref (https://www.crossref.org/community/ror/, https://www.crossref.org/blog/publishers-are-you-ready-to-ror/). Также эти идентификаторы использует ORCID (https://info.orcid.org/what-is-up-with-orcid-and-ror/) наряду с идентификаторами других платформ, таких как идентификаторы GRID, LEI, Crossref funder ID и Ringgold. Таким образом, каждая из указанных платформ указывает для каждой организации не только собственный идентификатор, но и идентификаторы «дружественных» платформ.

К сожалению, чаще всего речь идет о простом дублировании информации об одной и той же организации дружественными платформами. Наши эксперименты показали, что информация о российских научных организациях в ROR является копией информации об этих же организациях из базы данных GRID. При этом копируется как корректная информация, так и ошибочная. Например, устаревшая информация о веб-сайте исчезнувшего Новосибирского гуманитарного института в GRID (https://grid.ac/institutes/grid.445355.6) повторяется в ROR (https://ror.org/00nnwpb90).

Более важным примером скопированной ошибки является связь между страницей GRID, посвященной Сибирскому отделению Российской Академии Наук (Siberian Branch of the Russian Academy of Sciences,

(https://www.grid.ac/institutes/grid.415877.8) и идентификатором этой организации в ROR https://ror.org/02frkq021, которая выдает два «эквивалентных» названия этой организации: «SB RAS, ИНСТИТУТ КОСМОФИЗИЧЕСКИХ ИССЛЕДОВАНИЙ И АЭРОНОМИИ ИМ. Ю.Г.ШАФЕРА СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК».

Также, оба этих ресурса отличаются неполнотой данных, например, ни в одном из них нет информации про ИСИ СО РАН, как и про многие другие российские научные организации.

Именно примеры подобного рода указывают на необходимость интеграции информации между международными источниками данных и российскими ресурсами.

1. Сравнение информации в GRID и Wikipedia

Наряду с уже упоминавшейся базой данных eLIBRARY.ru, следует отметить такой источник данных как Wikipedia. База данных GRID поддерживает ссылки на страницы российских организаций в *англоязычной* Wikipedia, хотя было бы более естественно искать информацию о русскоязычных организациях в *русскоязычной* версии Wikipedia.

Поэтому первый эксперимент состоял в установлении соответствия между страницами GRID, а также англоязычной и русскоязычной версиями Wikipedia при помощи Wikipedia_API. Первоначально были проверены имеющиеся в GRID ссылки на англоязычную версию Wikipedia. Из 2019 российских организаций, имеющихся в GRID, только 412 имели ссылки на странички в англоязычной Wikipedia, и только 398 страниц англоязычной Wikipedia имели межязыковые ссылки (interlanguage links) на страницы русскоязычной Wikipedia. Для выяснения того, имеются ли в англоязычной и русскоязычной Wikipedia другие страницы русскоязычных организаций, описанных в GRID, осуществлялся поиск по таким атрибутам как URL веб-сайта организации, названиям организации и др. К сожалению, поиск в Wikipedia по адресу веб-сайта организации оказался не очень эффективным, и в конечном итоге пришлось осуществлять поиск по различным версиям названия организации. Поиск осуществлялся как по англоязычному названию, так и по русскоязычному названию в обеих языковых версиях Wikipedia. В результате этого поиска было дополнительно обнаружено 674 страницы в русскоязычной Wikipedia, соответствующие организациям из GRID, из которых 353 страницы были связаны межязыковыми ссылками с англоязычной Wikipedia. Всего же было обнаружено 835 страниц Wikipedia, соответствующих русскоязычным организациям, представленным в GRID. Таким образом, этот эксперимент показал, что в русскоязычной версии Wikipedia действительно имеется больше информации о российских научных организациях чем в англоязычной, но эта информация остается

недоступной для англоязычных баз данных из-за их ориентированности на англоязычную версию Wikipedia.

Основная проблема состояла в том, что явных ссылок не очень много, а поиск по названиям организаций сильно затруднен тем, что в разных базах данных приводятся разные названия одних и тех же организаций. Существующий, очень несовершенный алгоритм сопоставления планируется доработать.

2. Сравнение информации в GRID и eLIBRARY

На основе сопоставления количества данных о российских организациях, имеющихся в eLIBRARY.ru и GRID, возникает вопрос, за счет чего формируется такая значительная разница в количестве научных организаций. При внимательном рассмотрении было обнаружено, что eLIBRARY.ru действительно содержит очень большой список российских организаций, но далеко не все организации из имеющегося списка относятся к научным. В список организаций входят все федеральные министерства и организации, подчиненные этим министерствам, региональные ведомства, также коммерческие организации, банки, больницы, индивидуальные предприниматели, и пр. Например, в eLIBRARY.ru можно обнаружить такую организацию, как «Домостроительный комбинат № 7», за которой не числится ни одной публикации, и не имеется никакой информации кроме почтового и юридического адреса. В целом в списке организаций eLIBRARY.ru примерно одна треть (4505 организаций) вообще не имеет никаких публикаций, а еще примерно тысяча организаций имеет по одной публикации и столько же – по две публикации.

В этой базе данных также имеется немало организаций, прекративших свое существование. Например, у многих институтов СО РАН в качестве ведомства, которому подчиняется данная организация, указано Федеральное Агентство Научных Организаций (ФАНО). Страничка ФАНО имеется и базе GRID, но в русскоязычной Wikipedia имеется информация об упразднении этой организации 15 мая 2018 года. Эта же информация продублирована в наборе данных Wikidata (Federal Agency for Scientific Organizations , Q16711297).

Заметим, что алгоритм сопоставления данных из eLIBRARY.ru и GRID на основе названий организаций обнаружил эквиваленты только для половины российских организаций, указанных в GRID. Предполагается, что такой не высокий процент связан, прежде всего с реальной разницей в информации об организациях. В GRID, как и в eLIBRARY.ru, обнаружились организации, прекратившие свое существование, а также и заведомо ошибочные данные об организациях. С другой стороны, алгоритм сопоставления данных нуждается в улучшении. Наконец, возникает естественное желание использовать в качестве посредника между

русскоязычными и англоязычными источниками данных ресурс, имеющий доступ к разноязычным источникам данных.

3. Сопоставление данных eLIBRARY.ru и Wikidata

Примером очень перспективного источника данных является набор данных Wikidata. Wikidata возникла в 2014 году [6] как источник структурированных данных, для управления фактами в различных языковых версиях Wikipedia. Разработчики планируют сделать ее центральной платформой управления Wikipedia, интегрируя данные из всех языковых «глав» Wikipedia. Интеграция осуществляется присваиванием не зависящего от конкретной языковой версии идентификатора каждому объекту реального мира и объединения всех высказываний о заданном объекте реального мира из всех языковых версий Wikipedia. Помимо интеграции информации из различных языковых версий Wikipedia, этот набор данных так же, как и GRID поддерживает ссылки на глобальные библиографические ресурсы, при помощи указания идентификаторов в таких источниках данных как VIAF (Virtual International Authority File database https://viaf.org), Библиотеки конгресса США (Library of Congress, https://loc.gov/) и национальной библиотеке Германии (GND ID, https://portal.dnb.de). Также, помимо официального названия организации, в наборе имеется короткое название на русском и английском языках, информация о географическом положении (страна, область), даты возникновения и исчезновения сущности. Несмотря на эти явные достоинства, в наборе данных имеются и недостатки. Например, на страничке, посвященной СО РАН (https://www.wikidata.org/wiki/Q3032414) приводится тот же тот же самый неточный список подчиненных институтов, что и в базе данных GRID. В нем есть образовательные организации других ведомств, но зато не хватает некоторых институтов, реально подчиняющихся СО РАН. Например, в списке институтов, подчиненных СО РАН, нет ИСИ СО РАН, зато ИСИ СО РАН представлен как институт РАН, и на его страничке (https://www.wikidata.org/wiki/Q4201722) почему-то указано, что институт назван не в честь Академика А.П. Ершова, известного своими работами по информатике, а в честь Александры Петровны Ершовой, (Alexandra Petrovna Ershova, Q60830445) – российского театрального педагога. В данной работе рассматривается установление соответствия между русскоязычной eLIBRARY.ru и wikidata. В последующем эксперимент будет расширен данными из GRID.

Идея состоит в том, чтобы обогатить информацию об организациях, присутствующих в eLIBRARY.ru, полезной информацией из Wikidata, такой как ссылки на глобальные источники библиографических данных, время создания организации и др. Заметим, что Wikidata является источником структурированных данных и допускает извлечение информации при помощи запросов SPARQL. Но текущая неполнота данных не позволяет в

настоящий момент получить всю достоверную информацию при помощи запроса SPARQL. Так, например, запрос SPARQL, требующий выдать все российские научные организации, выдает только те организации, у которых в явном виде указано, что они расположены в России, а если сделать требование российской принадлежности необязательным (OPTIONAL), начинают выдаваться лишние организации, у которых вообще не указана страна принадлежности. Поэтому приходится решать задачу сопоставления сущностей программно.

Первым шагом для этого является идентификация идентичных организаций в двух источниках данных. В случаях, когда имеется информация об URL страницы организации в обоих источниках данных, задача является достаточной простой. Если полностью совпадают URL страницы организации, то осуществляется дополнительное сравнение названий организаций, и при коэффициенте совпадения пары названий из двух источников, превышающем 0.68, считается, что речь идет действительно об одной и той же организации, после чего из базы данных wikidata выбираются все интересующие нас значения атрибутов. К сожалению, ситуации, в которых оба источника данных имеют URL страницы организации, являются не такими уж частыми. Поэтому приходится прибегать к более трудоемким и менее надежным методам идентификации, где за основу принимаются другие атрибуты, в первую очередь, все доступные названия организации. К каждому из доступных названий применяется преобразование, переводящее все литеры в нижний регистр, затем из названий удаляются кавычки, убираются инициалы и фамилии, а также слова ООО, ОАО, АО, ЗАО, им., имени, г., города, СО, РАН, РАМН, ГУП, ДВО, АНО, открытое, акционерное, общество. Полученные таким образом строковые значения используются для поиска в Wikidata. Если в результате поиска по преобразованному названию обнаружена страница, из этой страницы извлекается основное название сущности, а также все доступные названия этой же страницы. В Wikidata эта информация содержится в поле «Также известный как»).

Также извлекается уникальный идентификатор сущности, URL страницы, страна и местоположение штаб-квартиры. Поскольку одно и то же название может соответствовать сущностям разного типа (например, организация и персона), происходит фильтрация по типу сущности и оставляются только организации. Затем осуществляется сравнение названий, для этого подсчитывается коэффициент совпадения названий организаций. Количество совпадающих слов в названии из Elibrary и Wikidata делится на количество слов в названии из Wikidata. Названия считаю совпадающими, если коэффициент > 0.68.

Далее, анализируется информация о стране организации, а также о местонахождении штаб-квартиры. Если в качестве страны организации указана не Россия, поиск останавливается, в базу записывается, что

организация находится не в России. Если Россия или страна не указана на странице Wikidata, извлекается информация о местонахождении штаб-квартиры — в Wikidata чаще всего указывается город, также извлекается информация про административно-территориальное образование. Заметим, что в Wikidata вместо Москвы, может быть указана, например, Московская область, что может привести к неправильному результату сравнения. Для решения этой проблемы используется специальный файл иерархии российских географических объектов. Таким образом, в настоящий момент организация считается верно идентифицированной в двух случаях:

А) Совпадают URL сайтов и некоторые пары вариантов названий.

Б) Сущность является организацией, названия организации в двух источниках совпадают, информация о сайтах не полна, организация находится в России.

Для всех организаций, распознанных алгоритмом как идентичные, информация об этих организациях объединяется на основе расширения онтологии schema.org.

На данный момент установлено соответствие между 3143 организациями Wikidata и eLibrary.ru. Получившийся экспериментальный источник данных будет в дальнейшем расширяться и интегрироваться с другими источниками данных, такими как GRID.

4. Заключение

Разработан метод интеграции из разноязычных источников данных. Создана экспериментальная версия базы данных научных организаций, состоящая из 3143 российских научных организаций, что в полтора раза превышает количество организаций, представленных в международных базах данных таких как GRID, но существенно меньше, чем базе данных eLIBRARY.ru. Планируется превратить эту базу в открытый и расширяемый граф знаний. Тем не менее, авторы считают, что для поддержания полноты и корректности информации о научных организациях, каждая научная организация должна поддерживать страницу своей организации с указанием всех идентификаторов организации в международных платформах.

Благодарности

Автор выражает благодарность Юрьевой И.О. и Чвыровой О.С., принимавшим участие в реализации программы сопоставления сущностей в разноязычных источниках данных.

Литература

1.Apanovich Z. Matching of authors and publications in multilingual bibliographic knowledge bases // CEUR Workshop Proceedings. SSI 2019 -

Proceedings of the 21st Conference on Scientific Services and Internet. 2020. C. 26-37.
2. Hajra, A., Radevski, V., Tochtermann K., Author profile Enrichment for Cross-linking Digital Libraries // Research and Advanced Technology for Digital Libraries Springer International Publishing. Lecture Notes in Computer Science. 9316 —2015. — P. 124-136
3. Mannocci, A.; Osborne, F., Motta, E. Geographical trends in academic conferences: An analysis of authors' affiliations// Data Science, 2019. —2(1) P. 181–203.
4. Global Research Identifier Database https://www.grid.ac/
5. Научная электронная библиотека eLIBRARY.ru https://www.elibrary.ru/
6. Ismayilov A., Kontokostas D., Auer S., Lehmann J., Hellmann S., Wikidata through the Eyes of DBpedia http://www.semantic-web-journal.net/system/files/swj1462.pdf

References

1.Apanovich Z. Matching of authors and publications in multilingual bibliographic knowledge bases//CEUR Workshop Proceedings. SSI 2019 - Proceedings of the 21st Conference on Scientific Services and Internet. 2020. C. 26-37.
2. Hajra, A., Radevski, V., Tochtermann K., Author profile Enrichment for Cross-linking Digital Libraries // Research and Advanced Technology for Digital Libraries Springer International Publishing. Lecture Notes in Computer Science. 9316 —2015. — P. 124-136
3. Mannocci, A.; Osborne, F., Motta, E. Geographical trends in academic conferences: An analysis of authors' affiliations// Data Science, 2019. — 2(1) P. 181–203.
4. Global Research Identifier Database https://www.grid.ac/
5. Научная электронная библиотека eLIBRARY.ru https://www.elibrary.ru/
6. Ismayilov A., Kontokostas D., Auer S., Lehmann J., Hellmann S., Wikidata through the Eyes of DBpedia http://www.semantic-web-journal.net/system/files/swj1462.pdf