DOI: 10.19181/4m.2022.31.1-2.2
EDN: GRRLBQ

Д.В. Мальцева, В.А. Ващенко, Л.В. Капустина
(*Москва*)

МЕТОДОЛОГИЯ ОБРАБОТКИ БИБЛИОГРАФИЧЕСКИХ ДАННЫХ НА РУССКОМ ЯЗЫКЕ ДЛЯ ПОСТРОЕНИЯ СЕТЕЙ КОЛЛАБОРАЦИИ
(на примере базы данных eLibrary)

Представлена методология обработки библиографических данных на русском языке на примере анализа публикаций российских социологов в электронной библиотеке eLibrary. Разработанный методологический подход подразумевает использование и адаптацию технологических решений для формирования базы библиографических данных, построения сетей для дальнейшего анализа и применения методов сетевого анализа для изучения различных областей знания. Описаны основные шаги сбора и предобработки данных на русском языке из eLibrary. На примере массива социологических публикаций из eLibrary рассмотрены типовые проблемы, возникающие на этапе предобработки библиографической информации об именах и аффилиациях авторов, предложены пути их решения. Разработанная методология позволяет сформировать базу библиографических данных и построить на ее основе сети коллаборации

Дарья Васильевна Мальцева – кандидат социологических наук, заведующая Международной лабораторией прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: dmaltseva@hse.ru
Василиса Андреевна Ващенко – стажер-исследователь Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: vvashchenko@hse.ru
Лика Владимировна Капустина – стажер-исследователь Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: lkapustina@hse.ru

для дальнейшего анализа. Методология применима для анализа различных публикаций русскоязычных авторов, представленных в eLibrary.

Ключевые слова: библиометрический анализ, библиографические сети, данные на русском языке, методология, социологическое сообщество, сети коллаборации, eLibrary

Введение

Современные исследования в области социологии исходят из важности изучения социального взаимодействия между учеными и их коллективами для определения их эффективности. В мировой практике научное взаимодействие, социальная и когнитивная структура различных научных областей успешно изучаются с помощью библиометрического анализа и анализа библиографических сетей – соавторства, цитирования, социтирования и библиографического сочленения между авторами и их коллективами [1; 2; 3].

В области наукометрического и библиометрического анализа предложен ряд методологических подходов и инструментов, варьирующихся по исследовательским задачам, принципам работы и степени знакомства пользователей с методологией сетевого анализа, выступающей общей рамкой для изучения библиографических сетей (например, программы VOSviewer и CitNetExplorer, пакеты Bibliometrix для R и Python, веб-приложение Biblioshiny). Источниками данных для этих инструментов выступают библиографические описания научных статей на английском языке, представленные в базах Web of Science (WoS), Scopus, Google Scholar и др. Заложенные в этих инструментах алгоритмы предобработки данных (дизамбигуации имен, лемматизации, токенизации слов и т.д.) ориентированы на англоязычные коллекции словарей и не могут быть напрямую использованы для нормализации данных

на других языках. При анализе данных на русском языке возникает задача по адаптации существующих или разработке новых подходов и инструментов библиометрического анализа. Примеры использования библиометрического анализа для изучения научных дисциплин в отечественной практике являются единичными [4; 5; 6], и полноценной методологии по сбору, предобработке и анализу библиографических данных на русском языке до настоящего времени не представлено.

Данная статья направлена на описание методологии работы с библиографическими данными на русском языке, сформированной авторским коллективом проекта «Паттерны коллаборации в российском социологическом сообществе: структура научных школ и возможные точки роста»[1]. Методология апробирована на библиографических описаниях публикаций российских авторов из научной электронной библиотеки (НЭБ) eLibrary, выбор которой обосновывается ниже в тексте статьи. В качестве примера взяты публикации российских социологов. Социологическое сообщество интересно как объект исследования в связи с нелинейным характером его развития, а также исторически обусловленными особенностями развития социологии как дисциплины в России [7]. Поскольку содержательное изучение в исследовании ориентировано на сети взаимодействия российских социологов, заключительным этапом разработанной методологии являются рекомендации по созданию сетей коллаборации. Таким образом, применение предложенной методологии позволяет сформировать базу библиографических данных и построить на ее основе сети коллаборации для анализа.

[1] Проект выполнен в рамках гранта Российского научного фонда № 21–78-00077 по Соглашению от 27.07.2021. Руководитель проекта – Мальцева Дарья Васильевна. Авторы выражают признательность за активное обсуждение проекта сотрудникам МЛ ПСА НИУ ВШЭ, Анушке Ферлигой, Владимиру Батагелю, Станиславу Моисееву, а также анонимному рецензенту за комментарии к статье.

Первый раздел статьи описывает некоторые примеры применения библиометрического анализа для изучения различных научных областей. Второй раздел знакомит читателя с основными шагами реализации методологии библиометрического анализа в целом. В третьем разделе представлена методология работы с библиографическими данными на русском языке: детально описан рабочий процесс сбора данных на площадке eLibrary и их последующей предобработки, включая основные проблемы и рекомендации по их устранению, а также детали построения сетей коллаборации. Статья завершается дискуссией и основными выводами.

1. Применение библиометрического анализа для изучения научных областей

При изучении научной коллаборации сети соавторства используются для анализа различных научных дисциплин и направлений в целом (наукометрика и инфометрика [8; 9], анализ социальных сетей [10; 11]) или в пределах национальных научных систем (например, изучались информационные науки в Аргентине [12], экономические науки в Польше [13]). На основе данных о соавторстве изучаются паттерны коллаборации в различных науках и проводится сравнение трендов развития научных дисциплин (биология, физика и математика [14; 15], математика и нейронаука [16], все исследовательские дисциплины в Словении [17; 18]). Сети соавторства изучены на многонациональном [19] и международном [20] уровнях. Сети соавторства используются для изучения трендов развития социологии в США [21; 22], США и Франции [23], Словении [24].

Хотя в отечественной периодике появляются публикации, где библиометрический анализ применяется для изучения научных областей на международном уровне [25; 26], использование этой методологии для изучения трендов развития науки в России не

распространено. Анализ социтирования (совместного цитирования двух публикаций третьей публикацией) в российских научных журналах использован для изучения этнологии и социологии [4; 5]. Единственный проект, где библиографические данные использованы для анализа российского социологического сообщества, посвящен изучению группы петербургских социологов [6].

2. Методология библиометрического анализа

В данной работе в качестве отправной точки для разработки методологии работы с данными на русском языке выступает подход к анализу библиографических сетей, разработанный В. Батагелем и его коллегами [27], уже использовавшийся для анализа научных сетей в таких областях, как сетевой анализ [11], кластеринг и блокмоделинг [27], научное рецензирование [28], наукометрия [9]. В отечественной литературе эта методология представлена при описании алгоритмического подхода к отбору источников для обзора литературы [25] и использовалась для изучения развития социологии [26].

Рабочий процесс подхода подразумевает использование технологических решений для: 1) формирования базы библиографических данных, 2) построения сетей для дальнейшего анализа и 3) применения методов сетевого анализа для определения трендов изучаемых областей. Остановимся на ключевых аспектах этих шагов.

2.1. Формирование базы библиографических данных

В оригинальной методологии [27] в качестве источника информации используется база WoS. Отбор единиц анализа происходит через выгрузку текстовых библиографических описаний работ по определенной тематике. В этих описаниях информация по каждой публикации (автор, название, выходные данные, цитируемая литература) размещена в соответствующем поле. Наличие

структуры в файле делает возможным извлечение нужной для анализа информации. Отобранные единицы анализа составляют массив библиографических описаний работ по исследуемому направлению.

2.2. Построение сетей

В рамках методологического подхода авторов создана специальная программа WoS2Pajek [27], которая позволяет выделять из массива информацию о связях элементов. Так как ключевым библиографическим элементом является публикация, фиксируются связи каждой работы с ее авторами, журналом, ключевыми словами и цитируемыми публикациями. На основе этой информации строятся одно- и двумодальные сети формата .net и файлы с дополнительной информацией. Получаемые сети связаны через набор узлов-публикаций и могут быть объединены через перемножение. Публикации имеют различное количество соответствующих библиографических единиц (авторов, ключевых слов, цитируемых работ), и при перемножении сетей некоторые единицы приобретают больший вес, поэтому важно проводить нормализацию их вклада [29].

2.3. Применение методов сетевого анализа

Полученные сети анализируются с помощью методов сетевого анализа в программе Pajek [30] для изучения предмета исследования на макро-, мезо- и микроуровнях анализа, что дает возможность изучения общей структуры, важных подгрупп и узлов в сети.

3. Разработка методологии библиометрического анализа для данных на русском языке

Рабочий процесс нашего подхода подразумевает те же шаги, что и оригинальная методология [27], описанная в разделе 2,

но требует специальных инструментов и техник для их адаптации. Основной упор сделан на формировании и обработке базы библиографических описаний на русском языке, из которой создаются файлы для сетевого анализа.

3.1. Формирование базы библиографических данных

Выбор источника данных. В начале исследования стоял вопрос о выборе репрезентативного источника библиографических данных для изучения отечественной науки, который был произведен после сравнения международной базы WoS и российских электронных библиотек КиберЛенинка и eLibrary. Каждый вариант обладал конкурентными преимуществами и ограничениями.

Выбор WoS позволял использовать оригинальную методологию, но с необходимостью ее адаптации для русского языка. Публикации российских авторов в WoS могут быть найдены в базах WoS Core Collection (WoS CC) и Russian Science Citation Index (RSCI/РИНЦ). WoS CC содержит публикации из международных журналов, коллекция RSCI – публикации российских авторов в российских и зарубежных журналах; у двух баз есть пересечение. Поскольку условия включения журналов в WoS CC являются более строгими по сравнению с другими базами, при использовании этой базы количество работ российских социологов будет значительно ниже реального объема. Технически поиск осложнен тем, что при указании русского языка для вывода публикаций требуется указать также тему; в выдачу не попадают статьи, написанные российскими авторами на других языках. При использовании WoS RSCI отбор публикаций также осложнен неоптимальной фильтрацией.

Коллекция РИНЦ представлена в eLibrary, где каждый ученый получает уникальный идентификатор (РИНЦ Author ID), что позволяет решать проблему дизамбигуации авторов. Возможны различные варианты формирования массива данных (поиск по тематике публикаций и журналов или по авторам).

Библиотека КиберЛенинка как альтернатива eLibrary позволяла проводить поиск по категории «социологические науки», но не давала возможности указать нужный временной период.

В связи с отсутствием технических ограничений, наличием опции идентификации авторов, вариативностью поиска публикаций, возможностью официальной закупки данных, было принято решение остановиться на базе eLibrary как на источнике данных исследования.

Стратегия формирования массива данных. Рассматривались различные варианты формирования массива публикаций на площадке eLibrary.

1. Отбор релевантных журналов и включение в массив их публикаций: поиск по журналам по тематике «Социология» привел к 802 журналам, многие из которых были международными. Отбор подходящих журналов на русском языке возможен в разделе сравнения журналов: через поиск по показателю журнала в рейтинге SCIENCE INDEX по тематике «Социология» находится 117 ведущих российских журналов.

2. Отбор авторов из изучаемой предметной области и включение в массив их публикаций: поиск по авторам, позволяющий сначала выбрать страну (Россия), а затем тематику («Социология»), привел к выдаче 8396 авторов.

3. Отбор публикаций по предметной области: поиск по тематике «Социология» вывел 782 046 публикаций за 2010–2019 гг., где часть статей написана зарубежными авторами (т.к. РИНЦ выборочно индексирует статьи в некоторых зарубежных научных журналах), а часть – опубликована в журналах, не входящих в РИНЦ.

Было принято решение остановиться на третьей стратегии и выбрать в качестве единицы анализа публикации, поскольку их принадлежность к определенной тематике строго зафиксирована (авторы могут делать вклад в разные научные области, а журналы могут включать различные дисциплины). Мы остановились на научных статьях, опубликованных в научных журналах,

индексируемых в РИНЦ. Из всех работ, относящихся по ГРНТИ к области «Социология», были отфильтрованы статьи, где по крайней мере одним из авторов является российский ученый (страна «Россия»). Интересующий нас период относился к 2010–2021 гг. Был составлен список из 75 232 ID публикаций, соответствующих условиям поиска, который использовался для сбора библиографических описаний.

Процесс сбора данных. Сбор данных проводился через API-сервис НЭБ. Выдача данных представляет собой XML-страницу структурированного вида с идентифицированными полями. Данные из нужных полей записывались в файл .csv с помощью языка программирования Python.

Запросы и обработка XML-страниц осуществлялись с использованием библиотек requests, xmltodict, xml, urllib; в процессе работы использовались библиотеки pandas, numpy, tqdm, time для облегчения работы с данными. По каждому ID статьи осуществлялся запрос к API с помощью requests и затем осуществлялась обработка XML-разметки страницы с помощью xmltodict, xml.etree.ElementTree. Полученная информация обрабатывалась по единому набору команд, извлекающих информацию о статье. Отдельно отметим большое количество конструкций try-except, использовавшихся для избежания системных ошибок, так как в некоторых статьях часть структурных элементов могла отсутствовать (аннотации, ключевые слова и т.д.). Некоторые проблемы возникали в связи с особенностями записи информации на XML-страницах. Каждая обработанная страница превращалась в строку pandas.DataFrame, содержащую информацию об одной статье, и присоединялась к массиву данных.

Несмотря на выбор «бережливого» подхода (с обработкой данных на каждой итерации и регулярным созданием локальных файлов-«чекпоинтов»)[1] и сбор данных с персонального ноут-

[1] Чекпоинт (checkpoint) – промежуточный файл, содержащий данные, загруженные к определенному моменту времени. Файлы-«чекпоинты» создаются

бука, средняя скорость выгрузки и обработки данных составила 2–3 итерации (статьи) в секунду. При средней скорости в 2,5 статьи в секунду для выгрузки и первичной обработки данных по 75 232 статьям потребовалось примерно 8 часов 20 минут чистой работы программного кода без учета других событий (отключение ноутбука из-за высокой нагрузки на процессор, возникновение технических ошибок в ходе исполнения цикла и трату времени на их устранение, проседание скорости WI-FI из-за большого числа пользователей).

Структура массива данных. С помощью механизма, описанного в предыдущем параграфе, по каждому ID публикации выгружались следующие данные о статьях.

1. Публикация: ID и название на русском и английском, DOI, дата, предметная область (второй уровень по ГРНТИ), количество страниц, число цитирований, аннотация и ключевые слова на русском и английском, поддержка, ссылка на оригинальную статью, краткое библиографическое описание.

2. Авторы: фамилии и инициалы авторов на русском и английском, РИНЦ Author ID и аффилиации и их ID.

3. Журналы: ID и название, ISSN и e-ISSN, импакт-фактор и включенность в ВАК, WoS, Scopus, РИНЦ и его ядро (RSCI), выпуск и номер, ID и название издательства.

Предобработка первичного массива данных

Типовые проблемы. Поскольку в базах публикаций библиографические описания представлены в полуструктурированном виде, первичный массив может содержать ошибки. Они возникают при ручном заполнении данных и требуют автоматической или ручной обработки. Проверка и предобработка первичного массива

раз в заданное пользователем количество времени или итераций цикла сбора данных. Их использование позволяет защититься от неожиданных сбоев в работе интернета или при выполнении кода, сохранив успешно загруженные данные, и продолжать работу с последней успешно выполненной итерации.

является важнейшим этапом библиометрического исследования, т.к. от качества базы зависят результаты.

Дизамбигуация единиц анализа (разрешение проблемы многозначности) – актуальная для библиометрических исследований задача. В первую очередь разрешение проблемы многозначности важно для имен ученых – авторов публикаций [31; 32]. Существует несколько видов неоднозначности имен авторов: одинаковые имена и фамилии у разных авторов, различия в написании или транслитерации одного и того же имени (или ошибки в написании), смена автором фамилии. Устранение неоднозначности направлено на то, чтобы найти все публикации автора и отличить их от публикаций других авторов с тем же именем. Классические подходы к дизамбигуации основаны на информации об авторах – аффилиации, контактах, годе публикации, соавторах, области исследований и т.п., которая может быть использована для обучения модели, выделяющей уникальных авторов [33]. Устранение неоднозначности имен рассматривается как проблема кластеризации – разделения документов на несколько кластеров, где каждый представляет автора [34, 35]; используются также сетевые подходы [36]. Проблему дизамбигуации можно решить и более простыми методами – при наличии у каждого автора уникального ID (SPIN-кода, РИНЦ AuthorID, ORCID или Researcher ID).

В eLibrary проблема многозначности имен авторов (в теории) решается наличием у каждого автора универсального ID, получаемого при регистрации в РИНЦ, и его аффилиацией с организацией, получающей ID при регистрации. Наличие такой информации по всем авторам могло существенно упростить процедуру обработки данных. Однако на этапе дескриптивного анализа первичного массива данных выяснилось, что в 75 232 публикациях только 19 739 авторов имеют уникальные ID (исключая значения 0 и *none*) и им соответствует 29 741 название и 1491 уникальное ID организаций. Еще 20% авторов и 32% аффилиаций имели значение ID, равное 0 или *none*. Проблема дизамбигуации была актуальной

для значительного числа единиц анализа. При работе с массивом были обнаружены проблемы, классифицированные как *проблемы недостатка имеющейся информации и проблемы неконсистентности имеющейся информации*, или проблемы *1-го* и *2-го* порядка.

Проблемы 1-го порядка связаны с отсутствием в базе нужной информации. На уровне авторов отсутствие ID связано с тем, что: 1) не все авторы в наборе данных зарегистрированы в РИНЦ и имеют ID, 2) у некоторых авторов, зарегистрированных в РИНЦ и имеющих ID, информация не всегда указана корректно и их ID либо отсутствуют, либо являются некорректными (приписаны ID других авторов или несуществующие). На уровне аффилиаций выявлены следующие проблемы: 1) у некоторых авторов (в том числе имеющих ID) не указана информация об аффилиации, 2) у части аффилиаций нет уникальных ID, так как они либо не зарегистрированы в РИНЦ, либо имеют некорректные названия, что не позволяет их идентифицировать. Из-за отсутствия информации в полях массива стоят значения 0 или none. Проблемы 1-го уровня должны решаться с помощью инструментов дизамбигуации, которые будут предложены ниже.

Проблемы 2-го порядка связаны с тем, что данные в массиве присутствуют, но не приведены к единому виду. Такие проблемы возникли при написании имен авторов и названий организаций. Особенно проблематичными становятся фамилии, содержащие мягкие знаки, букву «й» и двойные фамилии с дефисом. Ручное заполнение аффилиаций сопряжено со множественными опечатками и разнородными названиями для единого ID. Для решения проблемы дизамбигуации авторов нужно было сначала привести к единому формату наименования фамилий авторов и названий организаций (решить проблему 2-го порядка) и затем заполнить пропущенные в базе значения (решить проблему 1-го порядка).

Решение типовых проблем с именами авторов. В процессе дескриптивного анализа данных стало понятно, что данные о некоторых авторах подлежат автоматической или (в редких случаях)

ручной обработке. В табл. 1 описаны типовые проблемы, возникшие при обработке имен авторов, отсортированные по оценке их распространенности, а также предложенные пути их решения автоматизированными способами.

Некоторые данные об авторах приходилось восстанавливать вручную. При работе с массивом данных была проведена проверка на следующие факты и были предложены способы устранения ошибок.

1. Наличие знаков препинания в фамилиях на русском языке. В 7 случаях в фамилиях встречались вопросительные знаки (М?рат?ызы). Корректные фамилии восстановлены вручную, но при работе с массивами данных есть смысл проверки фамилий на наличие специальных символов.

2. Присутствие фамилий в колонках с инициалами на русском языке. Нашлось 108 статей, у авторов которых (преимущественно иностранного происхождения) в колонках с инициалами также были указаны фамилии. Мы не подобрали автоматизированного алгоритма разделения на фамилии и инициалы и восстанавливали данные вручную.

3. Наличие как минимум одного автора у каждой статьи. Выяснилось, что у 35 статей нет данных об авторах. Ручная сверка позволила восстановить часть информации, но при ограничении по времени «проблемные» наблюдения могут быть исключены.

В результате устранения типовых ошибок количество авторов значительно не изменилось, но приведение к единому формату позволило решать проблему дизамбигуации.

Решение типовых проблем с аффилиациями. При работе с аффилиациями были выявлены проблемы, затрудняющие использование присвоенных и создание новых ID.

1. Неоптимальное хранение данных о множественных аффилиациях в карточке публикации. Было обнаружено, что поле с аффилиациями хранится не как список, а как список со списком аффилиаций, что искусственно занижало размерность и

Таблица 1

ТИПОВЫЕ ПРОБЛЕМЫ, ВОЗНИКШИЕ
ПРИ ОБРАБОТКЕ ИМЕН АВТОРОВ

Столбцы	Инструмент детекции проблемы	Описание проблемы	Решение
Фамилии авторов на русском	Функция проверки языка	В ячейках с фамилиями авторов на русском – записи на английском	Транслитерация ячеек, где фамилии авторов должны быть записаны на русском, а записаны на английском, с английского на русский с использованием библиотеки transliterate для Python
Фамилии авторов на английском		В ячейках с фамилиями авторов на английском – записи на русском	Транслитерация ячеек, где фамилии авторов должны быть записаны на английском, а записаны на русском, с русского на английский с использованием библиотеки transliterate для Python
Инициалы авторов	Метод строк .isupper() в Python	Инициалы со строчной буквы	Привести все инициалы к прописным буквам и произвести транслитерацию соответствующих инициалов на английский

Окончание табл. 1

Столбцы	Инструмент детекции проблемы	Описание проблемы	Решение
Фамилии и инициалы на русском	Логическая проверка: входит ли в строку точка (в таблице с фамилиями)	Инициалы в ячейке фамилии на русском	Проверить все колонки с фамилиями авторов на русском языке на присутствие точек. Разделить фамилию и инициалы, перенести инициалы в соответствующую ячейку. Транслитерировать значения для ячеек с фамилией и инициалами на английском языке
Инициалы авторов на русском	Логическая проверка с использованием метода строк .count() Python	В инициалы на русском попадают полные имена или частичные инициалы	Использовать кастомную функцию, работающую с рядом проблемных случаев и приводящую строку с инициалами к виду «А.А.», транслитерировать значения на английский в колонку с инициалами на английском языке
Инициалы авторов на английском	Логическая проверка: начинаются ли инициалы российского автора на букву «Х»	При транслитерации буква «Х» транслитерируется в «Kh», и в инициалы попадает только «K»	В строках, где инициалы авторов на русском содержат букву «Х», – в ячейке с английскими инициалами замена на «Kh»

не позволяло идентифицировать и сохранять коды двойных и тройных аффилиаций на первой итерации. Проблема выявлена путем выборочной ручной проверки нулевых значений и решена посредством оптимизации кода.

2. Отсутствие названий на английском языке у большинства аффилиаций. Несмотря на наличие отдельного поля, данные между англоязычным и русскоязычным описанием аффилиаций не объединяются: у некоторых авторов аффилиации записаны на английском, а для русского языка в полях ID и названия указано "*none*".

3. Разнообразие названий аффилиаций и частичное отсутствие соответствующих ID. Одни и те же организации имеют разное написание – например, Московский государственный университет, помимо такой формы, может быть указан как «МГУ», «МГУ им. Ломоносова», «МГУ (им. Ломоносова)». В названии может встречаться город и статус организации (самостоятельная организация или филиал). При соответствии всем этим написаниям одного и того же ID, принадлежащего МГУ как зарегистрированной в eLibrary организации (2541), объединение единиц анализа под каноническим наименованием («МГУ») могло быть решено очень просто; проблема заключалась в том, что у части организаций ID от eLibrary отсутствовали. В этих случаях присвоение существующего ID по сходству названия было затруднительно и требовало предварительной обработки разных форм названий. Корректное автоматизированное преобразование могло бы быть осуществлено, если бы аффилиации записывались по единому порядку (например, указания о филиале располагались в конце названия). Но упорядоченность в структуре названий организаций отсутствует – указание на филиал встречается в начале и в конце строк, у филиала может быть собственное название или только указание на город расположения.

4. Наличие разноуровневых аффилиаций. Проблему ярче всего иллюстрируют аффилиации, связанные с крупными университетами, – может быть указан как головной университет, так и филиал

и/или отдельные исследовательские подразделения. В массиве данных нашлось 718 аффилиаций, связанных с РАНХиГС (сам институт и названия 47 его филиалов в разных формах), для НИУ ВШЭ – 404, для РАН – 308 аффилиаций. Ситуация характерна не только для крупных исследовательских центров, но и для крупных университетов Москвы и университетов федерального статуса. У части наименований даже при указании отдельного подразделения в качестве аффилиации все равно прикреплен корректный ID головной организации, но ряд наименований локального уровня таких связей не имеют и ID для них не прикреплен. Возникает необходимость присвоения таким аффилиациям названий головных организаций, а затем – соответствующих им ID, что возвращает к проблеме, описанной в п. 3.

Оптимизация кода сбора данных из XML-страниц позволила не зависеть от размерности поля, в котором хранятся аффилиации. Информация копировалась полностью и впоследствии преобразовывалась в список известной длины, с которым велась работа по нормализации аффилиаций.

Проблемы 3 и 4 решались единым образом. Колонки ID аффилиаций для английского и русского языков объединялись и сохранялись как список уникальных значений в двух колонках (см. табл. 2).

Таблица 2

ПРИМЕРЫ НОВЫХ ЗНАЧЕНИЙ
В ПРОМЕЖУТОЧНОЙ ОБРАБОТКЕ

Значение в колонке ID аффилиации на русском	Значение в колонке ID аффилиации на английском	Итоговое значение ID аффилиации
7113	7113	[7113]
‘none’	‘none’	[‘none’]
2541	‘none’	[2541, none]

Далее осуществлялся поиск по значениям с целью присуждения известных ID аффилиациям, его не имеющим. Были определены ключевые слова для часто встречающихся аффилиаций, позволяющие объединить под универсальным ID все подразделения организации (см. табл. 3). Корректность использования каждого кодового слова как критерия объединения проверялась экспериментально. Хорошими кодовыми словами стали фамилии деятелей, в чью честь названы вузы: тактика сработала для поиска незафиксированных аффилиаций, относящихся к МГУ им. Ломоносова, РГПУ им. Герцена, МГТУ им. Баумана и др.

Таблица 3

ПРИМЕРЫ КОДОВЫХ СЛОВ ДЛЯ КРУПНЫХ ОРГАНИЗАЦИЙ

Организация	РАНХиГС	НИУ ВШЭ	МГУ
Кодовые слова	«при президенте», «ранхигс», «ранх и гс»	«высшая школа экономики», «вше», «вшэ»	«ломоносов», «мгу»

Для объединения некоторых организаций без ID в единый кластер по общему признаку были созданы новые ID. Многие авторы указывали в качестве аффилиаций государственные органы и/или общеобразовательные учреждения, что позволило собрать названия аффилиаций в две группы: ‘Gos’ (государственные органы, например: Государственная Дума РФ) и ‘Sc’ (школы, например: МБОУ гимназия № 12 им. Г.Р. Державина). В итоге 171 уникальному автору в качестве первичной или вторичной была присуждена аффилиация ‘Sc’, а 32 уникальным авторам – аффилиация ‘Gos’.

Далее был создан словарь всех значений ID и возможных названий аффилиаций, встречающихся с этим ID. Для упрощения работы с аффилиациями в основной базе сохранены только их универсальные обозначения (например, все подразделения РАНХиГС сохраняются как РАНХиГС), однако в словаре значений сохранены

все написания аффилиаций, что позволяет при необходимости вернуться к уникальным названиям подразделений.

В результате обработки количество уникальных аффилиаций сократилось с 29 741 до 18 004, а уникальных ID – с 1491 до 1469 – за счёт нормализации и укрупнения ID.

Создание универсальных ID для авторов. Для решения проблемы дизамбигуации авторов были созданы универсальные ID, что позволило сократить количество повторов и оценить число уникальных авторов.

Новые ID для каждого автора содержали ключевую информацию: уникальный РИНЦ ID, инициалы, 8 символов фамилии на английском языке и ID всех аффилиаций автора и были построены по следующему принципу:

РИНЦ ID_Инициалы_Первые 8 символов фамилии_ID аффилиации

Отсутствующие аффилиации заполнялись строкой ‘none’. Множественные аффилиации были совмещены и включены в единый расширенный ID[1]. Аффилиации объединялись только в случае совпадения ID автора от eLibrary, не равного нулю: при уверенности в том, что данные относятся к одному и тому же автору. Примеры построенных ID приведены в табл. 4.

После создания ID была проведена процедура сокращения дублирующих ID, нацеленная на обнаружение опечаток, расхождений в аффилиациях и прочих деталей, позволяющих обнаружить упоминания автора как дубликат, а также обогатить новый ID информацией, содержащейся не в одном, а сразу во всех наименованиях автора.

Для этой задачи для всех возможных уникальных пар ID было рассчитано значение расстояния Дамерау-Левенштайна как количество различающихся символов между двумя строками [37].

[1] В итоговой базе с 37 790 уникальными авторами 94% авторов имеют одну аффилиацию. Несмотря на теоретическую необходимость фильтрации аффилиаций для их упорядочивания, такие кейсы в базе малочисленны, что позволяет временно опустить этот вопрос.

Таблица 4

ПРИМЕРЫ УНИВЕРСАЛЬНЫХ ID

РИНЦ ID	ID с отсутствующей или неопознанной аффилиацией	ID с одной идентифицированной аффилиацией	ID с несколькими идентифицированными аффилиациями
отсутствует	0_B_Guichard_none	0_I_Petrov_380	0_AI_Ivanov_351_875
присутствует	730489_MS_Val'des _none	429210_SI_Samygin_14461 865195_NG_Barinova_Sc	74486_SG_Maksimov_258_7082 137655_GE_ Zborovsk_290_1255_7366_14141

Таблица 5

ПРИМЕРЫ ИЗМЕНЕНИЙ, ВНЕСЕННЫХ В АВТОРСКИЕ ID НА ЭТАПЕ СРАВНЕНИЯ СТРОК

Тип проверки	Сравниваемые значения		Итоговый ID
Ручная	275901_VN_Bobkov_1432	275902_VN_Bobkov_1432	275902_VN_Bobkov_1432
Ручная	919176_RV_Parma_1074	919176_RV_Pyrma_1074	919176_RV_Pyrma_1074
Автоматическая	483923_OM_Shtompel_322	483923_O_Shtompel_322	483923_OM_Shtompel_322
Автоматическая	163439_VN_Petrov_1432	163439_VN_Petrov_210	163439_VN_Petrov_210_1432

Техника основана на методе сравнения строк для автоматического исправления орфографических ошибок, что близко к нашей задаче. В качестве потенциальных дубликатов рассматривались только пары, для которых значение метрики составило не более 4, так как 4 символа могут включать различие в аффилиации. Например, *12_II_Ivanov_1234* и *12_II_Ivanov_none* имели бы значение метрики, равное 4, что позволяет обнаружить эту пару как дубликаты и отказаться от '*none*' для обозначения аффилиации этого автора.

Сокращения дубликатов проводились только при наличии РИНЦ ID, поскольку иначе формальное основание объединения авторов отсутствовало. Были рассмотрены 8424 потенциальных пар-дубликатов. Из них 763 пары признаны дубликатами после нормализации инициалов, а 7146 – после нормализации аффилиаций, включая объединение нескольких аффилиаций для одного автора. После ручного анализа оставшихся строк на предмет выбора корректного написания фамилий была произведена еще одна итерация той же процедуры сравнения строк, чтобы убедиться в отсутствии дубликатов среди новых ID. На этом этапе были сомкнуты единичные ID за счет добавления дополнительных данных об аффилиации. В таблице 5 приведены примеры изменений, внесенных в авторские ID на этапе сравнения строк.

Изначальное количество авторских ID в базе данных было занижено ввиду невозможности опознать авторов, не имеющих РИНЦ ID (распределение по пропущенным значениям по очередности авторов представлено в табл. 6). С учетом обработки записей, где автор отмечен как отсутствующий, но на самом деле существует, удалось расширить число уникальных авторов в массиве (см. табл. 7).

Обработка новых ID позволила избавиться от 5581 дублирующейся записи авторов и сформировать наиболее приближенный к истинному список уникальных авторов. В результате все уникальные авторы были распознаны (не осталось значений "*none*" в их ID), и их число составило 37 790 авторов (см. табл. 7).

Таблица 6

ОЧЕРЕДНОСТЬ АВТОРА

Очередность автора	1	2	3	4	5	6	7	8
Доля ID, равных 0, %	18,9	23,1	22,5	26,2	26,9	33,7	24,8	21,5

Таблица 7

КОЛИЧЕСТВО УНИКАЛЬНЫХ ID АВТОРОВ

Этап предобработки	Количество уникальных ID авторов
До создания ID	19 741 (включая *none* и 0)
После создания ID	43 371
После обработки ID	37 790

Итоговый массив данных

Доступная база публикаций значительно трансформировалась в ходе преобразований с целью дизамбигуации авторов (см. табл. 8). По сравнению с изначальной базой, на 95% увеличилось количество уникальных авторов за счет идентификации тех, кто не имеет РИНЦ ID. Также были отфильтрованы и заменены «двойные» записи авторов, относящиеся к одному человеку, но содержащие расхождения в данных. Количество уникальных названий аффилиаций сократилось на 39,5% за счет нормализации описаний аффилиаций и приведения их к единому виду и создания универсальных описаний для аффилиаций с единым ID. Количество уникальных ID организаций сократилось на 1,5% – за счёт удаления некорректно заполненных ID.

Предобработка данных позволила справиться с наиболее важными проблемами, характерными для библиографических описаний статей в eLibrary, и сократить негативный эффект ручного заполнения данных для количественного анализа.

Таблица 8

СРАВНЕНИЕ ИЗНАЧАЛЬНОЙ И ФИНАЛЬНОЙ БАЗ ДАННЫХ

Элементы базы данных	Изначальная БД	Финальная БД
Уникальные ID авторов	19 366	37 790
Уникальные ID аффилиаций	1491	1469
Уникальные названия аффилиаций	29 741	18 004

3.2. Построение сетей и дальнейший анализ

После формирования итогового массива из соответствующих полей были выгружены данные для построения сетей коллаборации. Основой построения сетей соавторства выступает двумодальная сеть «Работа-Автор», где в первом наборе указываются все публикации, во втором наборе – все авторы, а далее фиксируются связи между ними. При построении сети необходимо проверить ее на отсутствие дублей и на то, что сила каждой связи равна 1. Путем перемножения и нормализации на основе этой сети строятся различные сети коллаборации [27].

Для создания сети с помощью Python сначала был создан файл формата .txt со списком всех узлов первого (публикации) и второго (авторы) наборов. В другом файле формата .txt были зафиксированы связи между узлами первого и второго наборов, извлеченные с помощью модулей pandas и numpy. Файлы были объединены вручную в единый файл, была добавлена информация об узлах и связях для прочтения программой Pajek (см. рис. 1). Готовый файл был сохранен в формате .net. Был сформирован файл с информацией о годах публикации отобранных работ, на основании которого сеть можно разделить на периоды для изучения коллаборации в динамике.

Могут быть построены другие двумодальные сети («Работа-Журнал», «Работа – ключевое слово») для построения сетей связей между авторами и ключевыми словами, ключевыми словами и журналами и т.д., а также дополнительные файлы с атрибутами

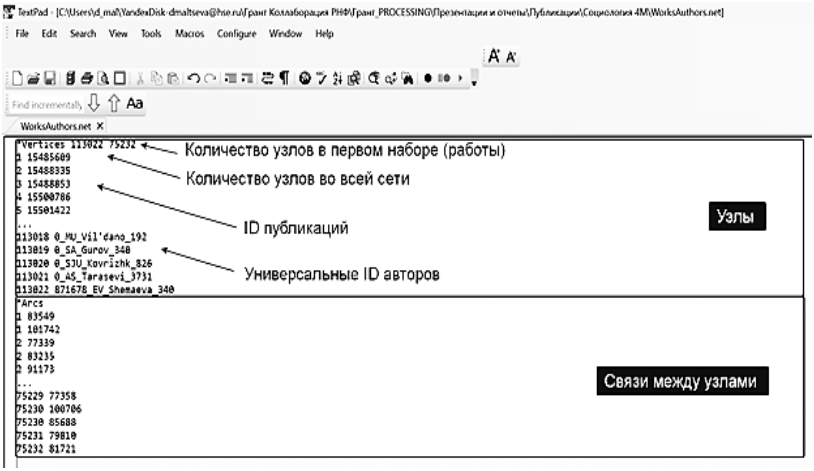

Рис. 1. Сеть «Работы-Авторы» в формате .net

узлов (количество страниц, цитирование). Получаемые сетевые данные можно использовать для решения разных исследовательских вопросов.

На следующем этапе был проведен дескриптивный анализ итоговой базы данных и сетевой анализ сетей коллабораций.

Дискуссия и основные выводы

Методологический подход к анализу библиографических данных включает этапы формирования базы, построения сетей и применения методов сетевого анализа для определения трендов. Авторы оригинальной методологии предложили ряд технологических решений для сбора, предобработки и анализа данных библиографических описаний на английском языке [27]. Если этап анализа данных предполагает использование программы Pajek, нечувствительной к языку записи данных, то этапы сбора и предобработки данных на русском языке требуют адаптации исследовательского инструментария. В оригинальной методологии база

формируется на основе данных из WoS, откуда полные библиографические описания скачиваются в полуструктурированном виде, и затем переводится в сетевые файлы с помощью программы WoS2Pajek. Для массива русскоязычных публикаций из WoS есть два ограничения: 1) база существенно занижает число публикаций, поскольку не индексирует большое число российских журналов; 2) заложенные в WoS2Pajek алгоритмы текстового анализа применимы только для английского языка.

Предлагаемый адаптированный методологический подход использует электронную библиотеку eLibrary как источник данных. Сбор осуществляется через API-сервис, где выдача данных представляет XML-страницу структурированного вида с идентифицированными полями, с использованием инструментов Python. В наиболее полном виде массив может включать всю информацию о публикации, ее авторах и журнале. Итоговые файлы сохраняются в форматах .csv и .xlsx.

Поскольку информация о публикациях представлена в базах данных в полуструктурированном виде, первичный массив может содержать ошибки, связанные с ручным заполнением данных по определенным полям. Мы классифицируем их как проблемы 1-го и 2-го порядка, или недостатка и неконсистентности имеющейся информации. Преодоление этих проблем важно для решения вопроса дизамбигуации авторов. Хотя преимущество eLibrary заключается в наличии у зарегистрированных авторов уникального ID, дескриптивный анализ показал, что проблема актуальна для большого числа единиц анализа. Мы описали типовые проблемы, возникающие на уровне авторов и аффилиаций, рассмотрели их причины и предложили решение автоматизированным и (в редких случаях) ручным способами. После устранения типовых ошибок на уровне авторов и аффилиаций были созданы универсальные ID для авторов и проведена процедура обогащения имеющихся и сокращения дублирующих ID, что решило проблему дизамбигуации авторов. В результате количество уникальных

авторов увеличилось на 95%. Предложенный подход позволяет справиться с наиболее важными проблемами библиографических описаний в eLibrary и сократить негативный эффект ручного заполнения данных об авторах и аффилиациях. ID авторов позволяют идентифицировать автора и организацию, что удобно для анализа и интерпретации данных.

Некоторые обозначенные проблемы связаны со спецификой eLibrary, но большинство выделенных проблем универсальны и могут возникнуть при работе с библиографическими данными на русском языке из других источников. При их обработке рекомендуется:

1) провести дескриптивный анализ и оценить долю пропущенных значений в полях с именами авторов, их аффилиациями и другими важными единицами анализа;

2) решить проблемы недостатка и неконсистентности имеющейся информации автоматическим и ручным способами:

a) для имен авторов:

– проверить наличие хотя бы одного автора у работы;

– проверить формат записей инициалов и фамилий, наличие лишних знаков, язык;

b) для названий организаций:

– проверить (выборочно) нулевые значения аффилиаций и убедиться, что они отсутствуют в базе;

– оценить распространенность множественных аффилиаций, проверить формат их записи;

– создать единый список ID организаций и на основе него присвоить ID организациям без ID;

– объединить разноуровневые названия организаций под ID головной организации (сохранив словарь соответствий);

– создать объединяющие названия для часто встречающихся категорий с единичными названиями (сохранив словарь соответствий);

3) решить проблему дизамбигуации авторов:

a) создать уникальные ID для авторов,

b) обогатить имеющиеся ID авторов дополнительной информацией,

c) провести сравнение и сократить дублирующие ID авторов.

Следующими этапами рабочего процесса является построение сетей и их анализ. Здесь можно вернуться к оригинальной методологии, построив двумодальные сети для построения изучаемых сетей. Для анализа используется Pajek; данные выгружаются из массива с помощью Python в виде текстовых файлов и сохраняются в нужных форматах.

Разработка комплексного методологического подхода для анализа библиографических данных на русском языке составляет научную новизну и практическую значимость проекта. Делая вклад в наукометрические и библиометрические исследования, использующие методологию анализа социальных сетей для построения сетей коллаборации, мы стремимся сделать этот вид анализа доступным для отечественного сообщества социальных исследователей. Алгоритмы сбора и предобработки данных опубликованы в открытом доступе на платформе GitHub[1] и могут использоваться для работы с данными eLibrary и других русскоязычных баз публикаций.

ЛИТЕРАТУРА

1. *Bar-Ilan J.* Informetrics at the beginning of the 21st century – A review // Journal of informetrics. 2008. Vol. 2, №. 1. P. 1–52. DOI: 10.1016/j.joi.2007.11.001. EDN: MISIBR.

2. *Mingers J., Leydesdorff L.* A review of theory and practice in scientometrics // European journal of operational research. 2015. Vol. 246, №. 1. P. 1–19. DOI: 10.1016/j.ejor.2015.04.002. EDN: UQPVRP.

3. *Rousseau R., Egghe L., Guns R.* Becoming metric-wise: A bibliometric guide for researchers / Ed. by W. Glänzel [et al]. Cambridge, MA: Chandos Publishing, 2018. 402 p. ISBN 9780081024744.

[1]Ссылка на страницу: https://github.com/Daria-Maltseva/Collaboration

4. *Сафонова М.А., Винер Б.Е.* Сетевой анализ социтирований этнологических публикаций в российских периодических изданиях: предварительные результаты // Социология: методология, методы, математическое моделирование (Социология: 4М). 2013. № 36. С. 140–176. EDN: RCFOWT.

5. *Винер Б.Е., Дивисенко К.С.* Когнитивная структура современной российской социологии по данным журнальных ссылок // Журнал социологии и социальной антропологии. 2012. № 15 (4). С. 144–166. EDN: PKOYXD.

6. Интеллектуальный ландшафт и социальная структура локального академического сообщества (случай петербургской социологии) / М.М. Соколов, М.А. Сафонова, К.С. Губа, Д.В. Димке; под ред. М.М. Соколова. М.: НИУ ВШЭ, 2012. 44 с. (Препринт / Высш. шк. экономики, Нац. исслед. ун-т; Серия WP 6, Гуманитарные исследования). EDN: QONWKT.

7. *Батыгин Г.С., Девятко И.Ф.* Социология и власть: эпизоды советской истории // Тоталитаризм и посттоталитаризм (Статьи и подготовительные материалы). Кн. 2. М.: ИС РАН, 1994. С. 174–201. ISBN 5-201-02478-5.

8. *Hou H., Kretschmer H., Liu Z.* The structure of scientific collaboration networks in Scientometrics // Scientometrics. 2008. № 75 (2). P. 189–202. DOI: 10.1007/s11192-007-1771-3. EDN: BMMNCG.

9. *Maltseva D., Batagelj V.* iMetrics: the development of the discipline with many names // Scientometrics. 2020. № 125. P. 313–359. DOI: 10.1007/s11192-020-03604-4. EDN: RHKXCV.

10. *Otte E., Rousseau R.* Social network analysis: a powerful strategy, also for the information sciences // Journal of information Science. 2002. Vol. 28, № 6. P. 441–453. DOI: 10.1177/016555150202800601. EDN: JNNEJB.

11. *Maltseva D., Batagelj V.* Collaboration Between Authors in the Field of Social Network Analysis // Scientometrics. 2022. № 6. P. 1–34. DOI: 10.1007/s11192-022-04364-z. EDN: VPCXGD.

12. A Global Comparison of Scientific Mobility and Collaboration According to National Scientific Capacities / Z. Chinchilla-Rodríguez, L. Miao, D. Murray [et al.] // Front. Res. Metr. Anal. 2018. P. 3–17. DOI: 10.3389/frma.2018.00017.

13. *Lopaciuk B.* Collaboration strategies for publishing articles in international journals – A study of Polish scientists in economics // Social Networks. 2016. Vol. 44. P. 50–63. DOI: 10.1016/j.socnet.2015.07.001.

14. *Newman P.* The structure of scientific collaboration networks // PNAS. 2001. Vol. 98, № 2. P. 404–409. DOI: 10.1073/pnas.98.2.404.

15. *Newman M.E.J.* Mixing patterns in networks // Physical Review E. 2003. № 2. P. 67. DOI: https://doi.org/10.1103/PhysRevE.67.026126.

16. *Albert R., Barabási A.-L.* Statistical Mechanics of Complex Networks // Reviews of Modern Physics. 2002. Vol. 74, № 1. P. 47–97. DOI: 10.1103/RevModPhys.74.47. EDN: LZWSIZ.

17. *Kronegger L., Ferligoj A., Doreian P.* On the Dynamics of National Scientific Systems // Quality & Quantity. 2011. Vol. 45, № 5. P. 989–1015. DOI: 10.1007/s11135-011-9484-3. EDN: SXTVAH.

18. Scientific collaboration dynamics in a national scientific system / A. Ferligoj, L. Kronegger, F. Mali [et al.] // Scientometrics. 2015. Vol. 104, № 3. P. 985–1012. DOI: 10.1007/s11192-015-1585-7. EDN: FAZSFM.

19. *Glänzel W., Schubert A.* Analysing Scientific Networks Through Co-Authorship // Handbook of Quantitative Science and Technology Research. Springer: Dordrecht, 2004. P. 257–276. ISBN 978-1-4020-2702-4. DOI: 10.1007/1-4020-2755-9_12.

20. *Wagner C.S., Leydesdorff L.* Network structure, self-organization, and the growth of international collaboration in science // Research policy. 2005. Vol. 34, № 10. P. 1608–1618. DOI: 10.1016/j.respol.2005.08.002.

21. *Moody J.* The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999 // American Sociological Review. 2004. Vol. 69, № 2. P. 213–238. DOI: 10.1177/000312240406900204.

22. *Hunter L., Leahey E.* Collaborative research in sociology: Trends and contributing factors // American Sociologist. 2008. № 39. P. 290–306. DOI: 10.1007/s12108-008-9042-1.

23. *Pontille D.* Authorship Practices and Institutional Contexts in Sociology: Elements for a Comparison of the United States and France // Science, Technology & Human Values. 2003. Vol. 28, № 2. P. 217–243. DOI: 10.1177/0162243902250905. EDN: JQALSF.

24. *Mali F., Ferligoj A., Kronegger L.* Co-authorship trends and collaboration patterns in the Slovenian sociological community // Corvinus journal of sociology and social policy. 2010. Vol. 1, № 2. P. 29–50. DOI: 10.14267/issn.2062-087X.

25. *Моисеев С.П., Мальцева Д.В.* Отбор источников для систематического обзора литературы: сравнение экспертного и алгоритмического подходов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2019. № 47. С. 7–43. EDN: MZXVXW.

26. *Булычева Е.Е., Мальцева Д.В.* Выделение актуальных тематик в социологии: взгляд сквозь призму анализа сети цитирований // Мониторинг общественного мнения: экономические и социальные перемены. 2020. № 6. С. 113–14. DOI: 10.14515/monitoring.2020.6.971. EDN: UGIDGS.

27. Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution / V. Batagelj, P. Doreian, A. Ferligoj, N. Kejžar. Chichester, West Sussex: John Wiley & Sons, 2014. 464 p. ISBN 1118915356, 9781118915356. DOI: 10.1002/9781118915370.

28. *Batagelj V., Ferligoj A., Squazzoni F.* The emergence of a field: a network analysis of research on peer review // Scientometrics. 2017. № 113. P. 503–532. DOI: https://doi.org/10.1007/s11192-017-2522-8.

29. *Batagelj V., Cerinšek M.* On bibliographic networks // Scientometrics. 2013. Vol. 96, № 3. P. 845–864. DOI: 10.1007/s11192-012-0940-1.

30. *Nooy W. de, Mrvar A., Batagelj V.* Exploratory social network analysis with Pajek. Revised and expanded edition for updated software. Cambridge; New York: Cambridge University Press, 2018. 420 p. ISBN 1108662099, 9781108662093. DOI: 10.1016/j.socnet.2005.12.002.

31. *Sanyal D.K., Bhowmick P.K., Das P.P.* A review of author name disambiguation techniques for the PubMed bibliographic database // Journal of Information Science. 2021. Vol. 47, № 2. P. 227–254. DOI: 10.1177/0165551519888605.

32. *Tekles A., Bornmann L.* Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches // Quantitative Science Studies. 2020. Vol. 1, № 4. P. 1510–1528. DOI: 10.1162/qss_a_00081.

33. *Treeratpituk P., Giles C.L.* Disambiguating authors in academic publications using random forests // Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. New York: Association for Computing Machinery, 2009. P. 39–48. DOI: 10.1145/1555400.1555408.

34. *Khabsa M., Treeratpituk P., Giles C.L.* Online person name disambiguation with constraints // JCDL '15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. New York: Association for Computing Machinery, 2015. P. 37–46. DOI: https://doi.org/10.1145/2756406.2756915.

35. A Unified Probabilistic Framework for Name Disambiguation in Digital Library / J. Tang, A.C.M. Fong, B. Wang, J. Zhang // IEEE Transactions on Knowledge and Data Engineering. 2012. Vol. 24, № 6. P. 975–987. DOI: 10.1109/TKDE.2011.13.

36. *Zhang B., Hasan M.A.* Name disambiguation in anonymized graphs using network embedding // Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017. P. 1239–1248. DOI: 10.1145/3132847.3132873.

37. *Damerau F.J.* A technique for computer detection and correction of spelling errors // Communications of the ACM. 1964. Vol. 7, № 3. P. 171–176.

Maltseva Daria V.,
Deputy head at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia, dmalceva@hse.ru

Vashchenko Vasilisa A.,
Research Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia, vvashchenko@hse.ru

Kapustina Lika V.,
Research Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia, lkapustina@hse.ru

Methodology of processing bibliographic data in Russian language to construct collaboration networks (using the example of the eLibrary database)

The methodology for processing bibliographic data in Russian is presented based on the example of analyzing publications of Russian sociologists in the eLibrary, Russia's largest electronic library of scientific publications . The developed methodological approach involves the use and adaptation of technological solutions to form a bibliographic database, builds networks for further analysis and applies network analysis methods to study various fields of knowledge. The main steps of collecting and preprocessing data in Russian from the eLibrary are described. Examining a corpus of sociological publications within the eLibrary, this study delves into common challenges encountered during the preprocessing stage of bibliographic information related to author names and affiliations. The paper suggests potential solutions to address these issues. Additionally, the paper suggests various solutions to address these challenges. The methodology is applicable to the analysis of various publications by Russian-speaking authors indexed in the eLibrary.
Keywords: bibliometric analysis, bibliographic networks, data in Russian, methodology, sociological community, network collaborations, eLibrary

References

1. Bar-Ilan J. Informetrics at the beginning of the 21st century – A review, *Journal of informetrics,* 2008, vol. 2, no. 1, p. 1–52.
2. Mingers J., Leydesdorff L. A review of theory and practice in scientometrics, *European journal of operational research*, 2015, vol. 246, no. 1, p. 1–19.

3. Rousseau R., Egghe L., Guns R. *Becoming metric-wise: A bibliometric guide for researchers.* Ed. by W. Glänzel [et al]. Cambridge, MA: Chandos Publishing, 2018. 402 p.
4. Safonova M.A., Viner B.E. Network analysis of co-citations of ethnological publications in Russian periodicals: preliminary results (in Russian), *Sotsiologiya 4M* (*Sociology: methodology, methods, mathematical modeling*), 2013, no. 36. p. 140–176.
5. Viner B.E., Divisenko K.S. Cognitive structure of modern Russian sociology based on journal references (in Russian), *Zhurnal Sotsiologii i Sotsialnoy Antropologii* (*the Journal of Sociology and Social Anthropology*), 2012, vol. 15, no. 4, p. 144–166.
6. Sokolov M.M., Safonova M.A., Guba K.S., Dimka D.V. *Intellectual landscape and social structure of the local academic community (the case of St. Petersburg sociology)* (in Russian). Ed. by Sokolov M.M. Moscow: HSE University, 2012, 44 p.
7. Batygin G.S., Devyatko I.F. Sociology and power: episodes of Soviet history (in Russian). *Totalitarianism and post-totalitarianism (Articles and preparatory materials)*, book 2. Moscow: IS RAS, 1994, p. 174–201.
8. Hou H., Kretschmer H., Liu Z. The structure of scientific collaboration networks in Scientometrics, *Scientometrics*, 2008, no. 75 (2), p. 189–202.
9. Maltseva D., Batagelj V. iMetrics: the development of the discipline with many names, *Scientometrics*, 2020, no. 125, p. 313–359.
10. Otte E., Rousseau R. Social network analysis: a powerful strategy, also for the information sciences, *Journal of information Science*, 2002, vol. 28, no. 6, p. 441–453.
11. Maltseva D., Batagelj V. Collaboration Between Authors in the Field of Social Network Analysis, *Scientometrics*, 2022, no. 6, p. 1–34.
12. Chinchilla-Rodríguez Z., Miao L., Murray D., Robinson-García N., Costas R., Sugimoto C.R. A global comparison of scientific mobility and collaboration according to national scientific capacities, *Frontiers in research metrics and analytics*, 2018, vol. 3, p. 3–17.
13. Lopaciuk B. Collaboration strategies for publishing articles in international journals – A study of Polish scientists in economics, *Social Networks*, 2016, vol. 44, p. 50–63.
14. Newman P. The structure of scientific collaboration networks, *PNAS*, 2001, vol. 98, no. 2, p. 404–409.

15. Newman M.E.J. Mixing patterns in networks, *Physical Review E*, 2003, vol. 2. p. 026126
16. Albert R., Barabási A.-L. Statistical Mechanics of Complex Networks, *Reviews of Modern Physics,* 2002, vol. 74, no. 1, p. 47–97.
17. Kronegger L., Ferligoj A., Doreian P. On the Dynamics of National Scientific Systems, *Quality & Quantity*, 2011, vol. 45, no. 5, p. 989–1015.
18. Ferligoj A., Kronegger L., Mali F., Snijders T. A., Doreian P. Scientific collaboration dynamics in a national scientific system, *Scientometrics*, 2015, vol. 104, no. 3, p. 985–1012.
19. Glänzel W., Schubert A. "Analysing Scientific Networks Through Co-Authorship", in: *Handbook of Quantitative Science and Technology Research*, ed. by Moed, H.F., Glänzel, W., Schmoch, U. Springer, Dordrecht, 2004, p. 257–276.
20. Wagner C.S., Leydesdorff L. Network structure, self-organization, and the growth of international collaboration in science, *Research Policy*, 2005, vol. 34, no. 10, p. 1608–1618.
21. Moody J. The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999, *American Sociological Review*, 2004, vol. 69, no. 2, p. 213–238.
22. Hunter L., Leahey E. Collaborative research in sociology: Trends and contributing factors, *American Sociologist*, 2008, vol. 39, p. 290–306.
23. Pontille D. Authorship Practices and Institutional Contexts in Sociology: Elements for a Comparison of the United States and France, Science, *Technology & Human Values*, 2003, vol. 28, no. 2, p. 217–243.
24. Mali F., Ferligoj A., Kronegger L. Co-authorship trends and collaboration patterns in the Slovenian sociological community, *Corvinus journal of sociology and social policy*, 2010, vol. 1, no. 2, p. 29–50.
25. Moiseev S.P., Maltseva D.V. Selecting sources for a systematic literature review: comparing expert and algorithmic approaches (in Russian), *Sotsiologiya 4M* (*Sociology: methodology, methods, mathematical modeling*), 2019, no. 47, p. 7–43.
26. Bulycheva E.E., Maltseva D.V. Highlighting Key Topics in Sociology: A Glance Through the Prism of Citation Network Analysis (in Russian), *Monitoring of Public Opinion: Economic and Social Changes*, 2020, no. 6, p. 113–140.
27. Batagelj V., Doreian P., Ferligoj A., Kejžar N. *Understanding large temporal networks and spatial networks: Exploration, pattern searching,*

visualization and network evolution. Chichester, West Sussex: John Wiley & Sons, 2014. 464 p.
28. Batagelj V., Ferligoj A., Squazzoni F. The emergence of a field: a network analysis of research on peer review, *Scientometrics*, 2017, vol. 113, p. 503–532.
29. Batagelj V., Cerinšek M. On bibliographic networks, *Scientometrics*, 2013, vol. 96, no. 3, p. 845–864.
30. Nooy W. de, Mrvar A., Batagelj V. *Exploratory social network analysis with Pajek. Revised and expanded edition for updated software*. Cambridge; New York: Cambridge University Press, 2018. 420 p.
31. Sanyal D.K., Bhowmick P.K., Das P.P. A review of author name disambiguation techniques for the PubMed bibliographic database, *Journal of Information Science*, 2021, vol. 47, no. 2, p. 227–254.
32. Tekles A., Bornmann L. Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches, *Quantitative Science Studies*, 2020, vol. 1, no. 4, p. 1510–1528.
33. Treeratpituk P., Giles C.L. Disambiguating authors in academic publications using random forests, *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, Singapore: ACM, 2009, p. 39–48.
34. Khabsa M., Treeratpituk P., Giles C.L. Online person name disambiguation with constraints, *JCDL '15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, Singapore: ACM. 2015, p. 37–46.
35. Tang J., Fong A.C.M., Wang B., Zhang J. A Unified Probabilistic Framework for Name Disambiguation in Digital Library, *IEEE Transactions on Knowledge and Data Engineering*, 2012, no. 24 (6), p. 975–987.
36. Zhang B., Hasan M.A. Name disambiguation in anonymized graphs using network embedding, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore: ACM, 2017, p. 1239–1248.
37. Damerau F.J. A technique for computer detection and correction of spelling errors, *Communications of the ACM*, 1964, vol. 7, no. 3, p. 171–176.