Дискуссии

УДК 001.12+303.2 DOI: 10.31431/1816-5524-2021-2-50-59-63

НУЖНА ЛИ УЧЕНОМУ ПЕРСОНАЛЬНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА?

© 2021 А.М. Асавин

Институт Геохимии и аналитической химии им. В.И. Вернадского РАН, Москва, Россия,111991; e-mail: aalex06@inbox.ru

Выполнен расчет объема литературы, который должен обработать ученый при проведении исследований в области науки о Земле. Примерная оценка обработанных ученым литературных источников составила 10000 работ за 20 лет. Для работы с таким большим объемом литературы необходимо использовать программы библиографических менеджеров. Рассмотрены пакеты программ EndNote и Mendelay. Обсуждается проблемы работы таких программных пакетов и их преимущества. Главным недостатком является отсутствие поддержки длительной истории работы с литературой. Кроме того поддерживается тематическая иерархия в атрибутитивных свойствах статей. Недостатки профессиональных пакетов заставили разработать собственный библиографический менеджер. Рассмотрены алгоритмы его работы, реализация в программе Microsioft Access. Предлагаемый алгоритм может быть использован и в других областях науки.

Ключевые слова: EndNote, Mendelay, библиография.

ВВЕДЕНИЕ

Как ни странно, ученым необходимо не только писать свои статьи, но и читать чужие. И надо отметить, что это занимает достаточно большой объем рабочего времени. Для сокращения этого времени в настоящее время активно используется различные технические решения. Часть таких средств рассмотрена в данной статье.

Информационные обеспечение исследований в области наук о Земле имеет несколько важных особенностей (?), что отличает эти работы от работ в других естественно-научных областях — физики, биологии, химии.

Во-первых, в своем понимании природных процессов исследователи пользуются не полной информацией. Объем ее определяется узким кругом природных объектов, с которыми они непосредственно работали. Им необходимо постоянно расширять круг рассматриваемых геологических объектов, поскольку найти закономерности геологических процессов и хорошо обосновать сделанные заключения можно только на достаточно больших выборках.

Во-вторых, природные объекты — структуры, месторождения, минеральные ассоциации и т.п. является по своей сути сочетанием случайных уникальных для данного места обстоятельств и общих закономерностей. В силу этого, каждый объект уникален и невозможно по их ограниченному числу точно выявить природные закономерности, отделить общие природные закономерности, от частных индивидуальных особенностей объекта.

В-третьих, при рассмотрении природных объектов приходится учитывать огромное количество параметров, которые относятся к разным областям науки — геофизики, геохимии, минералогии и др. При этом собственные области исследования автора могут быть достаточно далеки от этих областей.

Например, традиционно в публикациях в области наук о Земле присутствует раздел «Геологическая обстановка объекта исследования». Даже если, например, автор посвятил работу изучению состава флюидов в газово-жидких включениях из кварца гидротермальных жил, то он расскажет о вмещающих породах, тектонике района, возрасте проявления магматизма.

То есть исследователь будет вынужден включить в работу информационные источники, достаточно далекие от темы конкретного исследования. И это правильно, поскольку геологический объект является единой многопараметрической природной системой и, хотя в конкретном исследовании мы рассматриваем только одну из частных характеристик объекта, читатель должен хотя бы в первом приближении увидеть весь объект в целом.

Указанные объективные особенности исследований в области наук о Земле приводят к тому, что даже относительно небольшие статьи содержат очень длинные списки цитируемой литературы. Большой объем литературы, который должен обработать исследователь при подготовке статьи, наверняка в 2–3 раза больше окончательного списка, приведенного в статье, поскольку при поиске литературы по теме работы невозможно сразу найти работу отвечающую потребностям автора статьи. Назовем этот коэффициент — коэффициентом детальности изучения литературного обзора — Ko.

Естественно, по мере работы исследователя, его литературный багаж растет, и он отслеживает появившиеся новые публикации по теме, находит подтверждение своих заключений в ранее опубликованных работах классиков, в результате чего у него копится значительная по объему личная библиотека.

Попытаемся оценить объем этой библиотеки (Vb) исходя из следующих показателей: в год автор публикует (Na) около двух статей, в каждой из которых список литературы составляет (Va) порядка 100 работ, а время трудовой деятельности (t') порядка 20 лет.

$$Vb = Ko*Va*Na*t'$$

Итоговая сумма получается 8000 литературных источников. На самом деле в связи с тем, что в близких по темам работы списки литературы частично дублируются, Vb несколько меньше. Однако у исследователя очень часто наблюдается смена тематики работ, что приводит к резкому увеличению объема литературы, который он должен обработать.

Внесем индекс в параметр t'_1 обозначающий номер тематики t'_1, t'_2, t'_3.

Назовем параметр, означающий долю дублирующих в разных работах ссылок — Xdub. Например, предположим, что Xdub составит около 20%. Тогда расчетная формула выглядит так:

$$Vb = \sum_{t'=1}^{t'=3}(Ko*Va*Na)/Xdub \quad (1)$$

С учетом дополнительных двух-трех тем мы можем порядок суммарных цифр принять за 10000 источников.

Этот объем слишком велик, чтобы можно было удержать его в голове, что приводит к мысли о необходимости создания личной информационной системы. Собственно эта проблема и является предметом данной дискуссионной работы.

Второй вопрос — стоимость легального доступа к библиографическим информационным системам, о котором сильно переживает наше правительство и крупные издательства типа Elsevier, ScienceDirect и т.п. в таком объеме составит порядка 23 млн руб., то есть где-то стоимость неплохой московской квартиры.

Нужно ли это? Стоит ли тратить время, силы и деньги на создание личной информационной системы, а главное — на поддержку и эксплуатацию? Или следует выбросить эти мысли из головы и удовлетвориться наличием на собственном компьютере кучи файлов — статей, рассованных по различным директориям.

ПРОФЕССИОНАЛЬНЫЕ СИСТЕМЫ

Зададимся вопросом, существуют ли готовые профессиональные системы и насколько они удобны? В этом разделе мы рассмотрим два пакета EndNote и Mendeley. Это наиболее часто используемые англоязычными исследователями инструменты, хотя есть и другие интересные предложения (Перепелкин и др. 2019; Чебуков и др. 2017; MAXQDA 2020).

Пакет EndNote. Наиболее широко распространен библиографический менеджер EndNote. Его форматы данных (RIS) стали стандартом де-факто для других систем, да и идеология работы автора с базой данных в значительной степени дублируется в других пакетах. Большая часть работы переносится в Интернет среду, база данных состоит из ссылок с их атрибутикой для поиска данных. По тематике списки ссылок хранятся в виде отдельных файлов. При регистрации в большинстве современных библиографических Интернет-систем опционально для пользователя заводится личный кабинет — виртуальное пространство, в котором можно хранить выбранную для работы литературу. Физически там, конечно, хранятся только ссылки на литературу, но для пользователя это не важно. EndNote может подключаться к этим областям и, пользуясь своим интерфейсом, получать как полные тексты статей, так и библиографические списки. Далее, пользуясь собственным интерфейсом на локальной машине пользователя, проводится структурирование информации по коллекциям, ключевым словам и т.п.

Очень мощным средством является вывод выбранных в поиске работ в виде готовых библиографических списков под заданное форматирование, согласующееся с редакционными требованиями различных издательств. Есть масса готовых шаблонов и пользователь

может сам разрабатывать собственный шаблон. Не очень понятно, как реализовать последовательность в списке литературы по мере упоминания в статье, но, возможно, в пакете есть и такая возможность. Надо отметить, что для нелицензионных пользователей предельный объем записей в базе данных не велик — порядка 1000.

К главным преимуществам EndNote можно отнести возможность автоматически связываться с ведущими библиографическими базами данных Web of Science, Scopus, Berd. Это, конечно, относится к легальным пользователям, которые оплатили к ним доступ, или к исследователям тех институтов, у которых есть легальное подключение к этим системам.

При отсутствии таких условий пакетом можно пользоваться как дополнительной сетевой системой или для обмена данными.

Для поддержки связи ссылки с определенной тематикой в EndNote вводится понятие групп (Group). Число групп ограничено 5000. Важно, что ссылка может одновременно быть подключена к нескольким группам. Соответственно возможна сортировка по группам. Реализован совместный многопользовательский доступ к базе данных по библиографии на основе общего доступа к определенным группам.

Возможности импорта в базу данных включают списки литературы в формате Word, из отдельных pdf-файлов, и даже из сканированных документов. Так написано в руководстве. Но это достаточно сложная процедура, при которой в ручном режиме надо проверить правильность заполнения всех полей формата RIS. В базе данных осуществляется проверка на дублирование записей.

Пакет EndNote давно развивается и сейчас уже продается его 20-я версия.

Пакет Mendeley. Продается 19-ая версия, а в Мире сейчас насчитывается около 5 млн пользователей программы. Идеология интерфейса базы данных с пользователем очень похожа на предыдущий пакет. Вместо групп (Group), используются понятия коллекции (Collection). Нет возможности одновременного подключения работы к разным коллекциям. Появляется дополнительный признак теги (Tags). Это инструмент построения индивидуальной интеллектуальной сети тематик и привязке к ней библиографии.

В пакете более активно используется работа с файлами пользователя. Это с одной стороны хорошо, а с другой не очень. Программа автоматически меняет названия файла на составное имя — год, автор, название работы (в заданном порядке) и перемещает файлы в определенную директорию. Либо под каждого автора создается отдельная директория, куда помещаются все его работы. Все это интересно и полезно, но если у вас есть некий порядок размещения файлов по сложному дереву директорий, вы можете вдруг обнаружить, что все дерево поддиректорий уничтожено, а файлы переименованы. Причем поскольку программа одновременно делает копию базы данных в Интернете, то при синхронизации данных на вашем компьютере могут появиться удаленные ранее файлы, с другими именами или еще что-нибудь. В общем, программа достаточно самостоятельная и результат синхронизации (автоматический при запуске программы) не предсказуем. Однако сама идея ведения единой базы данных, не привязанной к компьютеру и синхронизируемой по запросу, очень привлекательна и вероятно стоит в этом разобраться.

Следует отметить, что и для EndNote и для Mendeley существуют большие сложности с кодировкой национальных языков. Примерно в половине случаев pdf-файлы с текстом на русском обрабатывается неадекватно, возникают сложности с кодировкой текста. При этом и для части китайских статей на английском возникает тот же эффект.

Интересна функция пакета, позволяющая разделить литературу на обработанную, или находящуюся в процессе чтения. Ведется трек чтения статьи, в который включаются замечания и маркировка текста. Интересно, насколько удобна эта функция?

Подводя итоги рассмотренным программам — «менеджерам библиографии», подчеркнем недостатки, которые является принципиальными для обеспечения управления базой данных. Пакеты не поддерживают тематическую иерархию для привязки ссылки. Что имеется в виду и почему это важно? В российской библиографии традиционно используется трехуровневая система УДК, поддерживающая классификацию по темам и подтемам. Иерархия помогает поддерживать смысловую связанность разных работ, создавать в некотором роде мнемоническую «нейронную сеть». Простое расположение фалов-ссылок в директориях и поддиректориях хорошо визуализирует этот процесс и позволяет упростить доступ к информации.

В профессиональных пакетах слабо проработан длительный процесс формирования библиографической базы, а ведь можно выделить, по крайне мере, три шага в этом процессе:

- составления списка предполагаемых для чтения работ;
- поиск работ и их чтение;
- оценка работ и включение в долговременное хранилище или маркировка ее как связанную с другой темой или как неинтересную для дальнейшего использования.

Это достаточно длительные процессы, а если еще одновременно проводится работа над несколькими статьями, то все значительно усложняется. В Mendeley эту проблему видят и стараются ее решить. Однако указанные проблемы заставили автора разработать собственную систему хранения библиографии.

Самодельная библиографическая система. После того как количество используемых статей превысило 1000 единиц, желание упорядочить процесс хранения превысило скорбь по потерянному времени на уборку файлового мусора и поиск информации на компьютере. В 2010–2017 гг. началась разработка собственной базы данных по библиографии. К настоящему времени удалось сделать более или менее законченную систему ввода и хранения библиографии.

Автор отнюдь не считает, что изложенные далее алгоритмы оптимальны для решения поставленных задач. Собственно, публикация данной заметки является приглашением к дискуссии по этим вопросам. Возможно, кто-то использует более эффективные алгоритмы обработки, имеет готовые решения или вообще не считает это существенной проблемой. Во всяком случае, я надеюсь, что изложенный далее опыт разработки будет интересен коллегам.

Надо отметить, что наибольшая трудность возникла при разработке алгоритма пополнения базы данных и согласованного хранения информации. Алгоритм должен быть достаточно простым, а с другой стороны, исключать дублирование статей и поддерживать смысловую привязку работ к определенной тематикеи исследователя. Кроме того, система должна быть толерантна к нестрогой форме ввода данных. Она должна поддерживать возможность постепенного упорядочения информации, а разработка должна быть максимально упрощена и требовать минимальных вложений сил. Поэтому при разработке использовались в основном возможности файловой системы, пакет Microsoft и ряд файловых менеджеров.

Вопрос тематической иерархии источников решен размещением файлов в четырех уровневой системе директорий. Система размещения поддиректорий строилась по принципу минимального времени доступа и содержания в директории нижнего уровня не более 30–50 файлов. В противном случае тематики детализировались на более высоких уровнях.

Вопрос постепенной обработки данных решен через дублирование дерева директорий основного хранилища — ядра библиографии и размещения кандидатов статей на размещение в этом ядре в дублирующей структуре. После прочтения или принятия решения об окончательной привязки статьи к определенной теме, небольшая программа обрабатывала дублирующую директорию и записывала название файла, его размер и имена поддиректорий верхнего и текущего уровня места хранения файла. Одновременно высчитывалась ссылка на авторов статьи и год издания. Перед окончательным перемещением файла по этой ссылке проверялась уникальность работы в базе. В случае неуникальности ссылки, но разного содержания работы к ссылке добавлялся буквенный индекс.

Таким образом, первой таблицей базы данных была таблица названия файла хранения, адреса его размещения (в терминах поддиректорий) уникального внутреннего номера в базе и размера файла.

Второй таблицей ядра базы было стандартное описание публикации (авторы, титул работы, журнал, издательство, страницы, и т.п.) с указанием на размер файла. Собственно связывание первой и второй таблиц проводилось по числу — размер файла. Здесь мы пользуемся довольно интересным свойством pdf-файлов – уникальностью размера каждого файла. Пользуясь этим свойством, проводился первичный поиск дубликата ссылок.

Для обеспечения ввода информации об обрабатываемых работах структура этих таблиц дублировалась в третью и четвертую таблицу, соответственно, и из них проводилось добавление информации после принятия окончательного решения о хранения информации в ядре. После добавления информации из этих таблиц происходило окончательное перемещение файла в место хранения.

Собственно эта структура обеспечивала основные задачи хранения и поиска информации. Остальные дополнительные таблицы в базе отвечали за форматированный вывод списков библиографии, временное хранение находящейся в поиске работы и ряд других служебных задач работы базы данных. В настоящий момент в базе находятся ~10000 записей и после отладки системы ввод новых 100 ссылок занимает примерно неделю. При этом основное время тратится непосредственно на чтение работ и их анализ. Изложенная примитивная схема была реализована в виде базы данных Access. Поскольку система индивидуального пользования, то не пришлось заботиться об экранных формах диалоговых окнах и т.п., что конечно не позволяет говорить о базе данных как о конечном продукте. Однако она выполняет свою задачу хранения и упорядочения библиографической информации и позволяет удобно работать с ней.

Подводя итоги, автор хотел бы открыть дискуссию на данную тему на страницах журнала и поделиться своим небольшим опытом

в обработке библиографической информации. Во всяком случае, автор будет благодарен коллегам за любые отклики по данной проблематике.

Работа выполнена в рамках темы по государственному заданию № 0137-2019-0014.

Список литературы [References]

Чебуков Д.Е., Изаак А.Д., Мисюрина О.Г. Управление списками публикаций авторов в библиографической информационной системе Math-Net.R // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18–23 сентября 2017 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2017. С. 456–460 [*Chebukov D.E. Izaak A.D., Misiurina O.G.* Upravlenie spiskami publikatsii avtorov v bibliograficheskoi informatsionnoi sisteme Math-Net R Nauchnyi servis v seti Internet trudy XIX Vserossiiskoi nauchnoi konferentsii 18-23 sentiabria 2017. Novorossiisk M: IPM im M.V. Keldysha, 2017. P. 456–460].

Перепелкин И.Н., Дунаев Р.А., Черняков А.Н. Средства автоматизации библиографической информации // Современные информационные технологии. 2019 № 3. Программное обеспечение, http://www.rusnauka.com/pdf/255078.pdf [*Perepelkin I.N. Dunaev R.A. Cherniakov A.N.* Sredstva avtomatizatsii bibliograficheskoi informatsii Sovremennye informatsionnye tekhnologii. 2019. № 3. Programmnoe obespechenie. http://www.rusnauka.com/pdf/255078.pdf].

MAXQDA Обзор Литературы в MAXQDA // VERBI Software. Consult. Sozialforschung. GmbH Berlin, Germany ww.maxqda.com 2020. 29 с. [MAXQDA Obzor Literatury v MAXQDA VERBI Software Consult Sozialforschung GmbH Berlin Germany ww maxqda com 2020. 29 p.].

DOES A SCIENTIST NEED A PERSONAL BIBLIOGRAPHY MANAGER?

A.M. Asavin

Institute of Geochemistry and Analytical Chemistry named after Vernadsky
Russian Academy of Sciences, Moscow, Russia, 111991

A calculation of the amount of literature that should be processed by a scientist when conducting research in the field of Earth science has been made. An approximate estimate of the volume of literary sources processed by a scientist is 10000 works over 20 years. To work with such a large volume of literature it is necessary to use programs of bibliographic managers. The EndNote and Mendelay software are reviewed. The problems of such software packages and their advantages are discussed. The main disadvantage is the lack of support for long-term literature review work. In addition, there is insufficient support for hierarchical systems of semantic-intelligent relationship of articles with problems of the science of the Earth. The limitations of professional software of the bibliography managers forced to develop our own bibliographic manager. Algorithms of its operating bibliographic system, which realized implementation in the Access Microsoft program, are considered. The proposed algorithm can be used in other fields of science as well.

Keywords: EndNote, Mendelay, bibliography.

Поступила в редакцию 20.05.2021 г.
После доработки 21.06.2021 г.
Принята в печать 28.06.2021 г.