Репозитории данных: теория и практика

Косычева Марина Александровна[1], Хорохорина Галина Анатольевна[2]

[1] ФГБОУ ВО «Московский государственный университет пищевых производств»
[2] ГБОУ ВО Московской области «Академия социального управления»

Корреспонденция, касающаяся этой статьи, должна быть адресована Косычевой М.А., ФГБОУ ВО «Московский государственный университет пищевых производств», адрес: 125080, город Москва, Волоколамское шоссе, дом 11. E-mail: kosychevama@mgupp.ru

Рассматривается необходимость создания и использования репозиториев данных для совместного и повторного использования данных исследователями, затрагиваются вопросы воспроизводимости исследований, увеличения вероятности цитирования. Приводятся критерии, которым должен соответствовать репозиторий. Анализируются аспекты, которые препятствуют распространению данных, среди них недоверие к данным, неправомерное использование данных другими исследователями.

Ключевые слова: репозитории данных, хранение данных, политика обмена исследовательскими данными, план управления данными

Глобальная цифровизация в современном мире не могла не затронуть область научных исследований. Скорость распространения и обмена информацией сегодня вынуждает научное сообщество внедрять политику обмена исследовательскими данными.

Под исследовательскими данными принято понимать количественную или качественную информацию, собранную ученым в ходе проведенного ими исследования. Исследовательские данные могут быть получены в результате экспериментов, наблюдений, моделирования, посредством опросов или интервью, или другими способами, или созданы на основе уже существующих данных. Данные исследований используются для подтверждения или обоснования результатов или выводов исследования. Их публикация и сохранение облегчают их повторное использование, их валидацию и способствуют воспроизводимости эксперимента (Wiley, 2018; Melero & Navarro-Molina, 2020). Говоря об исследовательских данных, подразумевают:

- первичные данные, которые были получены в результате исследования и отправлены авторами для последующей публикации;
- вторичные данные, которые были использованы авторами после анализа уже опубликованных данных;
- данные, которые были получены в результате эксперимента или наблюдений.

Политика открытого доступа к научным данным активно поддерживается Европейской комиссией и Европейским исследовательским советом (European Research Council (ECR))[1], основной принцип деятельности которых заключается в максимальной открытости исследовательских данных, за исключением тех данных, разглашение которых связано с этическими принципами.

Наиболее предпочтительным способом обмена исследовательскими данными на сегодняшний день считается загрузка наборов данных в онлайн репозитории данных и размещение ссылки на репозиторий в авторской статье (Зельдина, 2019).

Онлайн репозитории исследовательских данных – это крупные инфраструктуры баз данных, созданные для управления, совместного использования, доступа и архивирования наборов данных исследователей.[2] Репозитории могут быть специализированными и предназначенными для агрегирования дисциплинарных данных или более обобщенных данных в крупных областях знаний, таких как естественные или социальные науки. Онлайн-репозитории могут также объединять

[1] *Европейский исследовательский совет поддерживает передовые, междисциплинарные исследования и новаторские идеи в новейших областях научного знания с применением нетрадиционных и инновационных подходов. Миссия ERC состоит в том, чтобы поощрять исследования высочайшего качества в Европе за счет конкурентного финансирования и поддерживать ориентированные на ученых передовые исследования во всех областях научных изысканий на основе передового научного опыта. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/european-research-council*

[2] Brook, C. (December 5, 2018). What is a Data Repository? https://digitalguardian.com/blog/what-data-repository

Как цитировать

Косычева, М. А., & Хорохорина Г. А. (2020). Репозитории данных: теория и практика. *Health, Food & Biotechnology, 2*(1). https://doi.org/10.36107/hfb.2020.i1.s284

данные экспертов на глобальном или местном уровне, взаимовыгодно собирая исследовательские данные университета или консорциума университетов. Идея их создания заключается в том, что совместное использование данных позволяет ученым углубить свои исследования, получить более обоснованные выводы, проводить сравнение собственных данных и данных из репозиториев. Обмен данными и их совместное использование способствуют углублению исследовательских траекторий, появлению новых исследовательских трендов. Репозиторий позволяет исследовать, доказывать, проверять, способствовать прозрачности и подтверждать результаты исследователя другими экспертами за пределами опубликованной рецензируемой научной статьи. Размещение исследовательских данных в Интернете обеспечивает мгновенный доступ для группы исследователей, рассредоточенных по всему миру, для обмена, понимания и обобщения результатов их экспериментов. Кроме того, хранилища данных также предоставляют возможность ознакомиться с данными исследований, гипотезы которых не подтвердились. Тем самым, другие исследователи, ознакомившись с сутью «неудавшегося эксперимента», не тратят время на повторение ошибок, но получают возможность изначально строить исследование, избегая ложных посылов. Как следствие, позитивное отношение экспертов к открытому обмену данными посредством из размещения в репозиториях неуклонно растет.[3]

Репозитории данных делают возможным долгосрочное архивирование и сохранение данных путем приема / загрузки различных типов данных. Это могут быть как простые файлы Excel, SPSS, так и более специфические дисциплинарные форматы. В функционал репозитория как правило включена стратегия создания постоянных ссылок на данные для создания возможности их цитирования и мгновенного доступа. Иными словами, исследователи получают прямую ссылку на свои данные и вспомогательные файлы как для ее размещения в опубликованной статье. В случае использования этих данных другими исследователями, владелец данных получает цитирования. Обычно возможность цитирования обеспечивается с помощью идентификатора цифрового объекта (DOI) или универсального цифрового отпечатка (UNF), который позволяет впоследствии связывать данные и делает возможным взаимодействие и объединение архивов данных. В архивах данных также могут храниться пара-текстовые исследовательские материалы для последующего архивирования и обмена. Файлы данных могут включать электронные таблицы, полевые заметки, инструкции для лабораторий, мультимедийные материалы и специальные программы для анализа и работы с сопутствующими наборами данных.

Траектория инфраструктуры репозитория данных имеет свой жизненный цикл. Он начинается с эксперимента или исследовательского проекта и начального сбора данных, за которыми следуют загрузка, каталогизация, создание схемы дисциплинарных метаданных и присвоение DOI и / или UNF.[4] Репозитории позволяют реализовать мгновенный поиск, извлечение, связывание и загрузку данных. По мере развития хранилищ данных они позволят синтезировать наборы данных и поля данных, чтобы облегчить понимание, обнаружение и проверку.

Согласно политике обмена исследовательскими данными (Зельдина, 2019) авторы могут обращаться к двум ресурсам для поиска надежных репозиториев данных – это FAIRsharing.org и Repository Selector.[5] В них размещены репозитории, соответствующие ряду критериев, в том числе предоставления открытого доступа к опубликованным данным и обеспечения их долговременного хранения. Как правило, редакции журналов и издательств могут самостоятельно отбирать репозитории и включать их в список рекомендуемых. Крупные издательства, например Taylor & Francis, предлагают своим автором специальный сервис по руководству и размещению данных[6]. Springer Nature также рекомендует использовать проверенный список репозиториев в своем архиве[7]. В тех случаях, когда невозможно найти репозиторий для конкретного предмета, рекомендуется воспользоваться репозиториями общего назначения. Самыми распространенными считаются:

- 4TU.Datacentrum,
- ANDS contributing repositories,

[3] Uzwyshyn, R. (April, 2016). Research Data Repositories: The What, When, Why, and How. https://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml

[4] Там же

[5] *Repository Finder, пилотный проект Enhanced FAIR Data Project, возглавляемый Американским геофизическим союзом (AGU) в партнерстве с DataCite и сообществом ученых о Земле, космосе и окружающей среде, оказывает помощь в поиске подходящего репозитория для хранения исследовательских данных.* https://repositoryfinder.datacite.org/

[6] Data Repositories. https://authorservices.taylorandfrancis.com/data-sharing-policies/repositories/#

[7] Recommended Data Repositories. https://www.nature.com/sdata/policies/repositories

- Dryad Digital Repository,
- Figshare,
- Harvard Dataverse,
- Mendeley Data,
- Open Science Framework,
- Zenodo,
- Code Ocean.

Список критериев соответствия данных для их включения в репозитории данных должен быть прописан в политике журнала, равно как и механизм включения новых репозиториев в рекомендательный список.

Итак, с точки зрения соблюдения международных рекомендаций[8], выделены следующие критерии, которым должен соответствовать репозиторий, используемый авторами для хранения наборов данных:

- возможность загрузки наборов данных должна быть открыта для всех ученых, чьи исследования соответствуют тематике и техническим условиям репозитория, без каких-либо ограничений;
- репозиторий обязан предоставлять стабильный персональный идентификатор для всех загружаемых наборов данных (например, DOI);
- репозиторий должен предоставлять возможность распространять наборы данных под лицензией CC0 или CC BY (либо под лицензиями с аналогичными условиями) без ограничений на создание производных произведений или коммерческое использование;
- репозиторий должен бесплатно и без регистрации предоставлять доступ к наборам данных;
- репозиторий должен иметь долгосрочный план развития (включая финансирование), что служит гарантией сохранения наборов данных в будущем;
- репозиторий должен быть зарегистрирован в FAIRsharing.org;

репозиторий должен быть востребован в научном сообществе, о чем свидетельствуют размещенные наборы данных для большого количества опубликованных статей[9].

Как показывает международная практика государственные и частные исследовательские организации и исследовательские институты все чаще требуют от исследователей разработки планов управления данными (Data Management Plan). План управления данными предполагает описание жизненного цикла данных, собранных, обработанных и / или созданных во время проведения исследовательского проекта. Это документ, который идентифицирует и описывает такие вопросы, как процесс сбора данных, стандарты метаданных, используемые в их описании, и сохранение данных, а также отражает изменения или модификации, сделанные в ходе исследовательского проекта (Melero & Navarro-Molina, 2020). Таким образом, план управления данными предоставляет исчерпывающую информацию о данных и контексте, в котором они были созданы. Существует множество ресурсов для поддержки создания плана управления данными, например, для российских ученых Сибирское отделение РАН разработало вебнавигатор SciGuide, который помогает вести поиск качественных научных ресурсов[10,11].

Следует отметить, что существуют определенные факторы, которые могут препятствовать открытому обмену данными посредством репозиториев. К их числу относятся как технические факторы (Michener, 2015), так и человеческий фактор – сопротивление, конкуренция, привычки и т.д. (Fusi, Manzella, Louafi, & Welch, 2018). Проведенный в 2015 году опрос (Fecher, Friesike, Hebing, Linek, & Sauermann, 2015) выявил, что основным препятствием на пути к размещению данных в открытом доступе стала необходимость загрузить полученные в ходе исследования данные в репозиторий параллельно с направлением рукописи в редакцию журнала (то есть, до момента опубликования статьи). Авторы предпочитают сначала опубликовать статью, а потом делиться данными с научным сообществом. 80% респондентов выразили опасение, что другие исследователи могут опубликовать их данные в своих статьях раньше реальных авторов. Отсюда, очевидна необходимость разработки прозрачной процедуры совместного пользования данными. Исследователи считают, что, если бы у них была возможность решать, как и когда их данные будут повторно использоваться и кем, они предоставляли бы доступ к собственным данным значительно охотнее. Вторым препятствием выступает необходимость тратить усилия и время на размещение исследовательских данных в репозитории.

[8] THE FAIR DATA PRINCIPLES. https://www.force11.org/group/fairgroup/fairprinciples
[9] PLoS Repository Inclusion Criteria. https://journals.plos.org/plosone/s/data-availability#loc-repository-inclusioncriteria
[10] SciGuide. Научные ресурсы в открытом доступе. http://prometeus.nsc.ru/sciguide/page02.ssi
[11] Научные данные (Research data. Репозитории. Навигаторы). http://www.spsl.nsc.ru/resursy-gpntb-so-ran/big-data-repozitorii/

Важным инструментом соблюдения этических норм и авторских прав, по-прежнему остается цитирование. Существуют два основных подхода к отражению ссылок на наборы данных в статье: в отдельном разделе для списка наборов данных или в основном списке литературы. Отдельный раздел, включающий ссылки на наборы данных, более нагляден, однако включение ссылок в основной список литературы помогает приравнять ссылки на наборы данных к обычным ссылкам и упростить процесс их обработки (Cousijn et al., 2018). Ссылка на набор данных должна содержать следующую минимальную информацию (автор, год, название, DOI (или другой постоянный идентификатор), название репозитория.

Таким образом, репозитории заняли одно из ведущих мест в современной системе научной коммуникации, позволяя ученым обмениваться, распространять и использовать данные друг друга.

Литература

Зельдина, М.М. (2019). *Политика обмена исследовательскими данными (Data Sharing Policy).* Некоммерческое партнерство «Национальный электронно-информационный консорциум». https://dx.doi.org/10.24108/978-5-6040408-7-4

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M. & Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, *5*, 180259. https://doi.org/10.1038/sdata.2018.259

Fecher, B., Friesike, S., Hebing, M., Linek, S. & Sauermann, A. (2015). A reputation economy: results from an empirical survey on academic data sharing. *DIW Discussion Papers*, 1454. http://dx.doi.org/10.2139/ssrn.2568693

Fusi, F., Manzella, D., Louafi, S. & Welch, E. (2018). Building global genomics initiatives and enabling data sharing: insights from multiple case studies. *OMICS, 22*(4), 237–247. https://dx.doi.org/10.1089/omi.2017.0214

Melero, R. & Navarro-Molina, C. (2020). Researchers' attitudes and perceptions towards data sharing and data reuse in the field of Food Science and Technology. *Learned publishing, 33*(2), 163–179. https://doi.org/10.1002/leap.1287

Michener, W. K. (2015). Ecological data sharing. *Ecological informatics*, *29*, 33–44. http://dx.doi.org/10.1016/j.ecoinf.2015.06.010

Wiley, C. (2018). Data sharing and engineering faculty: an analysis of selected publications. *Science & Technology Libraries*, *37*(4), 409419. https://dx.doi.org/10.108%194262X.2018.1516596

Data repositories: theory and practice

Marina A. Kosycheva[1], Galina A. Khorokhorina[2]

[1] Moscow State University of Food Production
[2] Academy of Public Administration

Correspondence concerning this article should be addressed to Marina A. Kosycheva, Moscow State University of Food Production, 11 Volokolamskoe highway, Moscow, 125080, Russian Federation. E-mail: kosychevama@mgupp.ru

The need to create and use data repositories for sharing and reuse of data by researchers is considered, issues of reproducibility of research, increasing the likelihood of citation are discussed. Criteria the repository must meet are provided. The aspects that hinder the dissemination of data are analyzed, among them data mistrust and misuse of data by other researchers.

Keywords: data repositories, data storage, data sharing policy, data management plan

References

Zeldina, M.M. (2019). *Politika obmena issledovatel'skimi dannymi* [Data Sharing Policy]. Non-profit partnership "National electronic information consortium". https://dx.doi.org/10.24108/978-5-6040408-7-4

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M. & Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, *5*, 180259. https://doi.org/10.1038/sdata.2018.259

Fecher, B., Friesike, S., Hebing, M., Linek, S. & Sauermann, A. (2015). A reputation economy: results from an empirical survey on academic data sharing. *DIW Discussion Papers*, 1454. http://dx.doi.org/10.2139/ssrn.2568693

Fusi, F., Manzella, D., Louafi, S. & Welch, E. (2018). Building global genomics initiatives and enabling data sharing: insights from multiple case studies. *OMICS, 22*(4), 237–247. https://dx.doi.org/10.1089/omi.2017.0214

Melero, R. & Navarro-Molina, C. (2020). Researchers' attitudes and perceptions towards data sharing and data reuse in the field of Food Science and Technology. *Learned publishing, 33*(2), 163–179. https://doi.org/10.1002/leap.1287

Michener, W. K. (2015). Ecological data sharing. *Ecological informatics*, *29*, 33–44. http://dx.doi.org/10.1016/j.ecoinf.2015.06.010

Wiley, C. (2018). Data sharing and engineering faculty: an analysis of selected publications. *Science & Technology Libraries*, *37*(4), 409419. https://dx.doi.org/10.108%194262X.2018.1516596

How to Cite

Kosycheva, M. A., & Galina A. K. (2020). Data repositories: theory and practice. *Health, Food & Biotechnology, 2*(1). https://doi.org/10.36107/hfb.2020.i1.s284