К. А. Маслинский

О культуре работы с данными в филологии, или роль репозитория открытых данных

DOI: 10.53953/08696365_2024_189_5_358

Утверждение, что сегодня все филологи так или иначе работают с данными в цифровом формате, будет, пожалуй, сильно запоздалой констатацией. Однако если работа с источниками во всех филологических дисциплинах является важнейшей точкой приложения профессиональных стандартов и методологической рефлексии, то в отношении цифровых данных очень распространен сугубо инструментальный подход. Стоит сделать шаг от работы с цифровыми копиями бумажных документов к организации собственных цифровых данных — разного рода таблиц, текстовых коллекций и баз данных, как о профессиональных стандартах забывают. Исследователь остается наедине с известным ему набором цифровых инструментов, что нередко ведет к повторению распространенных ошибок и тиражированию неудачных решений.

Обсуждение форматов представления данных в цифровой форме, принципов их публикации, способов долгосрочного сохранения, инструментов и правил обращения с ними в гуманитарных областях остается уделом относительно узких профессиональных групп — специалистов, работающих в области цифровой гуманитаристики (digital humanities) или занятых количественными исследованиями. В этих областях сложился консенсус относительно того, что цифровые способы представления структурированных научных данных не являются методологически нейтральными, поскольку решения, принятые при оцифровке материалов, уже представляют собой реконструкцию историко-культурных процессов. Затем эти решения и в целом структура оцифрованных данных ложатся в основу научной аргументации[1]. Поэтому необходима дальнейшая профессионализация работы со структурированными цифровыми данными в филологических исследованиях. Для этого нужны как более широкая методологическая дискуссия для достижения консенсуса в разных дисциплинарных кругах, так и институциализация лучших практик и построение научной инфраструктуры.

Данная статья посвящена обсуждению опыта работы над одним инфраструктурным проектом, направленным на решение этой задачи. В 2020 г. на базе Лаборатории цифровых исследований литературы и фольклора в Институте русской литературы (Пушкинском Доме) РАН мы создали Репозиторий открытых данных по русской литературе и фольклору (https://dataverse.pushdom.ru). Репозиторий — это веб-ресурс для хранения и публикации научных данных, которые авторы предоставляют в свободный доступ другим исследователям. Инфраструктурная роль репозитория в том, чтобы повысить видимость и доступность данных, поддерживать культуру цитирования данных и в конечном итоге способствовать расширению количественных исследований в литературоведении и фольклористике. Одновременно с этим репозиторий может выступать как инструмент влияния на научную практику работы с данными. Редакция репозитория не просто пассивно ожидает

1 См.: *Володин А.Ю.* Данные в цифровых гуманитарных исследованиях // Цифровые гуманитарные исследования. Красноярск: Сибир. федер. ун-т, 2023. С. 21—38.

материалов от авторов, но работает над тем, чтобы способствовать публикации данных и делать это в соответствии с международными стандартами публикации открытых данных и нашим представлением о научной ответственности.

Задача статьи — с одной стороны, представить концепцию репозитория широкой научной общественности, а с другой — критически обсудить опыт работы редакции репозитория.

Миссия и задачи репозитория

Идея создания тематического репозитория научных данных для российской филологии обязана своим появлением концепции открытой науки. Это научно-реформаторское движение манифестирует ценностную парадигму, в которой открытость, доказательность и воспроизводимость заявлены как неотъемлемые идеалы научной работы[2]. Можно сказать, что открытая наука — это утопический проект, в том смысле, что он предлагает идеализированные образы социальной практики для научного сообщества. Репозитории открытых данных — один из таких утопических образов. Слово «утопический» может ассоциироваться с чем-то не существующим в действительности, однако в нашем случае речь идет об уже действующем научно-инфраструктурном проекте. Его утопичность — в тех целях, которые мы перед собой ставим. В целеполагании репозитория можно выделить два плана: ближний (операциональные задачи) и дальний (более широкая миссия).

Ближайшая задача — способствовать тому, чтобы в российском академическом сообществе в целом и, в частности, в его административном сегменте, публикация данных признавалась как полноценная научная публикация. Это важно для развития эмпирической базы цифровых и количественных исследований, поскольку на подготовку данных уходят огромные исследовательские ресурсы, а получить за это какое-то признание или минимальное соответствие критериям административной отчетности очень трудно. Необходима более четкая мотивация для исследователей, подталкивающая их инвестировать усилия в подготовку и публикацию данных. Институционализация публикации данных — один из очевидных путей. Для международного научного сообщества этот тезис уже стал общим местом. Шаги, предпринятые в этом направлении научным сообществом, включают разработку принципов цитирования данных[3], инфраструктуры для публикации и атрибуции данных (репозиториев, систем присвоения постоянных идентификаторов, таких как DOI), постепенное распространение требований к публикации данных одновременно с публикацией исследования[4] и параллельное распространение практики цитирования данных в собственных исследованиях. Наш репозиторий создан в том числе для приобщения к этим практикам российского филологического сообщества.

Долгосрочная цель в плане публикации данных — это мир, в котором публикация научных данных является нормой, а не исключением. На сегодняшний день

---

2 См.: *Fecher B., Friesike S.* Open science: One term, five schools of thought // Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing / Ed. by S. Bartling, S. Friesike. Cham: Springer International Publishing, 2014. P. 17—47. (Электронное издание.)

3 См.: *Altman M. et al.* An introduction to the joint principles for data citation // Bulletin of the Association for Information Science and Technology. 2015. Vol. 41. No. 3. P. 43—45.

4 См.: *Stodden V., Guo P., Ma Z.* Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals // PloS one. 2013. Vol. 8. No. 6. e67111. P. 1—8. https://doi.org/10.1371/journal.pone.0067111

в мировой науке это пока далеко не так, как показали исследования доступности данных в разных дисциплинах[5]. Нам хотелось бы, чтобы публикация данных воспринималась как вклад в эмпирическую базу науки как коллективного социального предприятия. Для академического сообщества это вполне закономерный подход. Публикация научных результатов уже давно представляет собой нечто само собой разумеющееся, но публикация данных — пока еще нет.

Другой аспект миссии репозитория — сделаться важным инфраструктурным узлом для отечественной филологии. Публичный веб-ресурс, на котором систематически помещаются качественные открытые данные по русской литературе и фольклору, имеет потенциал к тому, чтобы стать точкой входа для поиска эмпирического материала в российских литературоведении и фольклористике. Понятно, что нельзя собрать на одном централизованном ресурсе все существующие цифровые данные в нашей дисциплине. Но собрать многие важные данные как минимум о русской литературе мы, как кажется, могли бы. Этот аспект миссии репозитория прекрасно согласуется с более широкой миссией Института русской литературы, с которым аффилирован репозиторий. Институт создан и развивается как организация, которая отвечает за сохранение и популяризацию русского литературного наследия. Многие работы сотрудников института связаны со сбором, систематизацией и публикацией литературных данных. Со временем менялись только формы представления. В доцифровую эпоху доминировали книжные публикации (такие как академические собрания сочинений) и бумажные картотеки (самый известный пример — картотека Б.Л. Модзалевского). С наступлением компьютерной эры в дополнение к ним появились такие цифровые формы, как библиографические базы данных, доступные локально (библиографическая база данных «Пушкиниана», разработанная отделом пушкиноведения ИРЛИ) или в виде веб-интерфейса (электронный справочник «Источники русской агиографии»[6]). Задача репозитория — вывести практику публикации цифровых данных на современный мировой уровень.

Более амбициозная и долгосрочная цель — способствовать расширению количественных и корпусных исследований в литературоведении и фольклористике. Доступность ресурсов — инфраструктуры и данных — чрезвычайно важное условие для такого рода работы. Методология статистического анализа текстовых и историко-культурных феноменов за последние десятилетия шагнула далеко вперед. Однако, чтобы по-новому изучать историю литературы и эволюцию литературных форм, нужны по-новому представленные данные. В качестве аналогии можно вспомнить о том, как изменилась русистика с появлением Национального корпуса русского языка. Мы бы хотели, чтобы подобные трансформации произошли и в литературоведении, и репозиторий — шаг в этом направлении.

Наконец, самая долгосрочная составляющая миссии репозитория — архивация данных. Данные в интернете, электронные публикации в какой бы то ни было форме в большинстве своем исключительно эфемерны и исчезают за считанные годы[7].

---

5 См., например: *Culina A. et al.* Low availability of code in ecology: A call for urgent action // PLoS Biology. 2020. Vol. 18. No. 7. P. e3000763.

6 http://odrl.pushkinskijdom.ru/Default.aspx?tabid=1937

7 См.: *Goh D.H.-L., Ng P.K.* Link decay in leading information science journals // Journal of the American Society for Information Science and Technology. 2007. Vol. 58. No. 1. P. 15—24; *Zeng T., Shema A., Acuna D.E.* Dead science: Most resources linked in biomedical articles disappear in eight years // Information in contemporary society: 14th international conference, iConference 2019, Washington, DC, USA, March 31 April 3, 2019, Proceedings 14. Springer, 2019. P. 170—176.

Редкие ресурсы способны пережить одно-два десятилетия. В то же время сохранность данных — это ключевой компонент рецепта воспроизводимости результатов в эмпирической науке. Проблема долгосрочной сохранности цифровых научных данных хорошо осознана в мировой науке, в Европейском союзе существует ряд инфраструктурных проектов, ставящих задачи долгосрочной архивации данных[8]. В отечественной науке эти задачи обсуждаются слишком мало, и создатели репозитория хотели бы, чтобы он был образцом и в этом отношении.

Архитектура репозитория

Репозиториев научных данных в мире уже довольно много. Наш репозиторий имеет узкую дисциплинарную направленность — мы ограничили себя только русской литературой (словесностью) и фольклором. В этом отношении он радикально отличается от репозиториев общего назначения, таких как Zenodo (https://zenodo.org/) или гарвардский Dataverse (https://dataverse.harvard.edu/), больших университетских или национальных научных репозиториев. Специализированные репозитории для публикации данных определенной тематики на сегодняшний день более характерны для естественно-научных дисциплин, обрабатывающих большие объемы данных (астрономия, геология и т.п.), чем для дисциплин гуманитарных. Пожалуй, наиболее близок к нашему репозиторию по задачам и масштабу репозиторий открытых данных проекта, посвященного американской литературе и культуре после 1945 г., — Post45 Data Collective (https://data.post45.org/).

Репозиторий — конкретный веб-ресурс, для реализации которого необходима техническая платформа. Для этой цели был выбран Dataverse (https://dataverse.org/), свободное программное обеспечение с открытым кодом, разработанное для Гарвардского университета. Гарвардский университет предлагает два варианта использования: либо свободно публиковать данные в гарвардском репозитории, либо создавать собственные репозитории для публикации данных на базе других институций. Мы пошли по второму пути. При использовании Dataverse для нашего репозитория в кодовую базу проекта потребовалось внести два существенных изменения. Во-первых, перевести интерфейс на русский язык. Во-вторых, поменять схему регистрации DOI для публикуемых данных. В гарвардском репозитории предполагается, что его держатель — крупная институция, обычно богатый американский университет или европейское агентство, регистрируется у провайдера DOI и закупает номера крупными пакетами, например 10 тысяч в год. DOI регистрируется автоматически непосредственно в момент загрузки автором набора данных в репозиторий. Автор же может принимать решение о публикации данных. Мы изменили процедуру таким образом, что регистрация DOI происходит не при создании датасета (набора данных), а при его публикации, что может быть сделано вручную, когда редакция репозитория принимает решение о публикации данных. Это освобождает от необходимости закупать большие пакеты DOI.

Чтобы данные признавались публикацией, их нужно встроить в инфраструктуру цитирования. Удобство платформы Dataverse не только в присвоении DOI каждому набору данных, но и в том, что она размещает пример формата цитирования на центральном месте на странице данных, делая идею цитируемости этого материала весьма наглядной. Модель публикации и цитирования данных на плат-

8 См.: *Антокольский А.Б., Володин А.Ю.* Информационная инфраструктура цифровых гуманитарных исследований // Цифровые гуманитарные исследования. С. 244—263.

форме Dataverse состоит в том, что опубликованные данные зафиксированы и платформа дает гарантию, что их всегда можно будет получить в этой же форме. Это принципиально отличается от распространенной практики цитирования цифровых ресурсов с помощью ссылки на URL. Цифровизация научной работы привела к тому, что интернет-ресурсы часто воспринимаются как формы публикации, а интернет в целом — как универсальный цифровой архив. Но интернет — это не архив, а средство связи. То, что вы получили по определенному адресу URL сегодня, не обязано будет соответствовать тому, что вы получите по этому адресу завтра; более того, завтра вы можете вообще ничего не получить. Гарантия неизменности данных после публикации в репозитории сближает цифровую публикацию данных с традиционной напечатанной книгой. Платформа Dataverse предлагает возможность обновления набора данных, но при условии явного указания номера версии данных и сохранения доступа к предшествующим версиям.

Задача институционализации публикации данных в глазах научного сообщества (и администрации) требует развития доверия к данным. Для этого необходимо, чтобы ученые, никак не аффилированные с конкретным репозиторием и институтом, сформировали мнение относительно достоверности, качества и полезности данных. Условия доверия к формату публикации научных данных не являются новым изобретением. Можно привести в пример академические собрания сочинений, в отношении которых существует высокий уровень доверия к качеству проведенной текстологической работы. Аналогичным образом необходимо развивать институт доверия к цифровым данным, опубликованным в репозитории. Традиционный инструмент повышения доверия к качеству публикации — независимое рецензирование. Вклад репутации тех исследователей, которые берутся выступить рецензентами, поддерживает основания для доверия. В нашем репозитории принята процедура рецензирования данных до публикации, и его можно рассматривать как рецензируемое (peer-reviewed) онлайн-издание для данных. Другим основанием для укрепления доверия в научном сообществе выступает репутация ИРЛИ, высокие стандарты которого в работе с источниками мы стараемся переносить и в цифровую плоскость.

Как и для любого репозитория открытых научных данных, для нашего репозитория ориентиром в отношении стандартов подготовки и публикации данных являются принципы FAIR (findability, accessibility, interoperability, reusability — обнаружимость, доступность, совместимость, повторное использование)[9]. Следование этим принципам задает конкретные решения: открытая публикация данных (отсутствие ограничений в доступе), публикация данных под свободными лицензиями (право повторного использования), следование цифровым стандартам, избегание проприетарных форматов файлов. Однако приходится учитывать, что для повторного использования данных нужно выполнение не только технических, но и социальных условий. Важнейшее из последних — понятность данных. Поэтому в репозитории ИРЛИ все публикации сопровождаются развернутой документацией, которая помогает понять, какая именно цифровая модель действительности скрывается в каждом опубликованном наборе букв, чисел и файлов и каким может быть ее научное или общественное значение.

Важно отметить, что ответственность репозитория заканчивается в тот момент, когда пользователь скачивает файлы данных. Это значительно отличается от «цифровой платформы» в привычном понимании, характерном для многих облас-

9 См.: *Wilkinson M.D. et al.* The FAIR guiding principles for scientific data management and stewardship // Scientific data. 2016. Vol. 3. No. 1. P. 1—9.

тей цифровой гуманитаристики. Можно сказать, что репозиторий предлагает данные не как готовый продукт в красивой упаковке, а в виде набора «сделай сам».

Опубликованные данные

Основные тематические разделы репозитория во многом мотивированы типами задач, сформировавшими практики сбора и систематизации данных в филологии еще в доцифровую эпоху. Сейчас в репозитории четыре тематических раздела: корпусы текстов; библиографические данные, которые трактуются максимально широко: любые сведения об издании, упоминании или циркуляции произведений, не включающие текстов самих произведений; биографические данные; данные для воспроизводимых исследований.

За три года существования репозитория опубликовано в общей сложности 29 датасетов. В 18 из опубликованных датасетов (62%) авторами или соавторами выступили сотрудники, аффилиированные с Лабораторией цифровых исследований ИРЛИ.

Конечно, до создания централизованного репозитория, который охватывал бы все важные данные литературоведения, еще очень далеко. Но в тематике уже опубликованных данных можно проследить общие проблемы и исследовательские парадигмы, в рамках которых современные филологи считают нужным создавать цифровые данные.

Целая серия публикаций показывает, что для одной из важных тем современной теории литературы — проблемы литературного канона — цифровые данные являются чрезвычайно актуальным инструментом. Распространенный подход — систематически собранные данные об упоминаниях авторов и произведений в определенных типах источников за значительный период времени. Так, можно объединить в цикл три датасета, посвященные фиксации школьного литературного канона в общей сложности более чем за два столетия: литературные произведения, включенные в школьные хрестоматии (*Вдовин А., Лейбов Р., Казакова Е.* Хрестоматии Российской Империи с 1805 по 1912 г.); произведения и авторы в советских школьных программах по литературе (*Кондра М., Казакова Е.* Программы по литературе для средней школы с 1919 по 1991 г.) и в постсоветских образовательных стандартах (*Кокорин А.* Литературные произведения в государственных стандартах и программах для средней школы 1998—2022 гг.). Другой источник сведений о каноническом статусе произведений — включение в антологии, альманахи, сборники, чему посвящен еще один датасет, охватывающий данные за вторую половину XIX века (*Олещук А.* Лучшие образцы русской литературы (1849—1900): антологии избранной поэзии и прозы, литературные сборники и альманахи, сборники для легкого чтения, антологии для народа, антологии для женщин). Чрезвычайно редкие данные о рецепции литературного канона и устном бытовании литературных произведений представлены в датасете «Бытование литературных текстов в ГУЛАГе» (*Луговская Д.А.* и др.).

Когда исследовательская проблема формулируется таким образом, что для ее решения требуется охватить значительное количество произведений, принадлежащих определенному жанру, периоду или источнику, в нашу цифровую эпоху естественным инструментом исследователя становится формирование электронного корпуса текстов. Заметное место среди публикаций этой группы в нашем репозитории занимают коллекции поэтических текстов, посвященных отдельным жанрам: элегии (*Мартыненко А.* Корпус русских элегий 1815—1835 гг.), песне (*Шеля А.* Корпус «русской песни» 1800—1840-х гг.), балладе (*Иванова М.* Корпус русской

литературной баллады 1840 гг.). Источником для формирования корпуса может послужить и канонический автор, например таков датасет «Байрон в русских переводах» (*Бодрова А.* Байрон в русских переводах 1810—1860-х годов). Некоторые корпусы, созданные авторами под конкретную исследовательскую задачу, обладают широким потенциалом для применения и во многих других исследованиях. Примерами могут послужить «Корпус русской нарративной прозы XIX века», включающий 500 произведений как классических, так и малоизвестных авторов (*Собчук О., Лекаревич Е.* Корпус нарративной прозы XIX в.) и корпус публикаций журнала «Современник» (*Вожик Е.* Корпус публикаций журнала «Современник» (1847—1866)).

Значительная часть опубликованных датасетов посвящена детской литературе. Характер данных о детской литературе при этом весьма разнообразный. Самая крупная публикация в этой группе — корпус русской прозы для детей и юношества, охватывающий более 3000 художественных и нехудожественных произведений за 1900—2020 гг. (*Маслинский К., Лекаревич Е., Алейник Л.* Корпус русской прозы для детей и юношества). В силу ограничений, создаваемых авторским правом, этот датасет не содержит текстов произведений, однако в нем представлены полные метаданные и производные частотные данные, позволяющие воспроизвести большую часть вычислений по текстам корпуса, если они основаны на лексической частотности и сочетаемости. Другой тип материалов — библиографический: приведенная в машиночитаемую табличную форму оцифрованная библиография детской книги (1918—1984), основанная на 18 томах указателей И.И. Старцева и его последователей (*Маслинский К.* Библиография детской книги (1918—1984) и библиография детской книги русского зарубежья в Европе (1919—1954). Литературная и педагогическая критика — третий важный источник данных. В репозитории представлены как сведения об упоминаниях авторов и произведений в критических статьях и рецензиях 1860—1885 гг. (*Лучкина О.* Авторы и произведения для детского чтения в критике 1860—1880-х гг.), так и корпус критических статей о детской литературе русского зарубежья (*Димяненко А.* Критика детской литературы русского зарубежья в периодических изданиях 1920—1940-х гг.). Наконец, целый ряд публикаций содержит данные и код для воспроизведения результатов количественных исследований по материалам детской литературы. Это исследования об изменении семантичеких контекстов понятия «счастье» в детской литературе за сто лет (*Маслинский К.* Сто лет счастья в детской литературе (1920—2020): сталинский канон и его долгосрочные последствия); о гендерных различиях в домашнем труде литературных персонажей (*Лекаревич Е.* Домашние дела литературных персонажей); количественное исследование стилистики прозы Виктора Голявкина (*Маслинский К.* Стилистика детской прозы Виктора Голявкина: Синтаксический профиль); о жанровых различиях в упоминании животных в детской литературе (*Maslinsky K.* Replication Data for: How Exactly does Literary Content Depend on Genre? A Case Study of Animals in Children's Literature).

Совершенно другой подход к цифровой репрезентации данных о литературном процессе дает рассмотрение литературы через сеть социальных и литературных связей писателей с другими лицами. В доцифровых форматах этот подход в литературоведении манифестировался в таких источниках, как биографические словари и указатели имен к собраниям сочинений. Оцифровка таких источников позволяет применить к ним современные методы визуализации и сетевого анализа. Самые значительные публикации на эту тему в нашем репозитории посвящены литературе XVIII в. Это реконструированная на основе Словаря русских писателей XVIII в. сеть персоналий (*Орехов Б.* Словарь русских писателей XVIII века: сеть персоналий) и дополняющая ее и согласованная с ней сеть, описывающая русско-европейские

литературные связи того же периода (*Бакиров Р., Орехов Б.* Русско-европейские литературные связи XVIII века). Несколько иной тип данных, также построенных вокруг персоналий, — сведения о социальных и литературных связях конкретного писателя с другими лицами. Таковы данные о встречах Ходасевича, извлеченные из его «камер-фурьерского журнала» (*Орехов Б., Успенский П., Файнберг В.* «Камер-фурьерский журнал» В. Ходасевича), и датасет, построенный на основании указателя имен и названий из академического собрания сочинений Чехова (*Северина Е.М., Северин Н.Н., Петров К.О.* Указатель имен и названий: полное собрание писем А.П. Чехова).

Подавляющее большинство текстовых корпусов, опубликованных в репозитории, представляют собой неразмеченные текстовые материалы, где каждое произведение или издание сопровождается подробными метаданными, но структурные и содержательные элементы внутри текста не маркированы. Однако один датасет представляет собой публикацию корпуса с богатой разметкой: это текст «Войны и мира» с размеченными упоминаниями персонажей, включая принадлежность реплик и семантические роли персонажа в тексте (*Скоринкин Д.* Персонажи «Войны и мира» Л.Н. Толстого: вхождения в тексте, прямая речь и семантические роли). Такие данные могут быть чрезвычайно полезны для обучения и тестирования систем машинного обучения, ориентированных на автоматическую высокоуровневую разметку литературного текста, таких как BookNLP[10].

Среди датасетов, отражающих целые типы потенциально существующих и востребованных в нашей дисциплине цифровых данных, можно назвать датасет, позволяющий воспроизвести стилеметрическое исследование об авторстве «Тихого Дона» (*Орехов Б.* Стилеметрические данные «Тихого Дона» и современной ему прозы).

Наконец, упомянем блок наборов данных, посвященных творчеству А.С. Пушкина. Это прежде всего подготовленный сотрудниками Пушкинского Дома Индекс произведений и писем Пушкина. В этом индексе систематизированы и представлены в машиночитаемом виде все известные сведения о письменном наследии поэта, причем каждому произведению присвоены компактные уникальные идентификаторы (как это принято в работе с наследием крупных авторов прошлого, таких как Аристотель, Бах, Моцарт и др.). Помимо индекса, подготовлен к публикации корпус стихотворений Пушкина, в котором представлены тексты, выверенные по академическим собраниям произведений и подготовленные для машинной обработки (*Вожик Е.И., Казакова Е.О., Лисюков Р.А.* Корпус стихотворений А.С. Пушкина).

Некоторые из опубликованных датасетов связаны с более традиционными для цифровой гуманитаристики веб-ресурсами; они содержат те данные, на которых построен и которые представляет пользователю веб-ресурс. Так, датасет Корпус русской прозы для детей и юношества согласован с теми данными, с которыми можно ознакомиться с помощью поискового интерфеса (конкорданса), доступного на сайте: http://detcorpus.ru. Аналогичным образом среди веб-ресурсов Пушкинского Дома доступен поисковый интерфейс по нарративной прозе XIX в.[11] — 500 романов, представленных в датасете Корпус нарративной прозы XIX в. Другой пример — веб-приложение, позволяющее в интерактивном режиме рассматривать сеть взаимосвязей русских писателей XVIII в.[12], отражающее данные соответствующего датасета «Словарь русских писателей XVIII века: сеть персоналий».

10 https://github.com/booknlp/booknlp
11 http://corpora.pushdom.ru/
12 https://nevmenandr.github.io/rus-dict18-persons/

Редакционная политика и процедура публикации

Основные требования к данным, которые могут быть опубликованы в репозитории, обусловлены их потенциальной полезностью для аудитории репозитория, прежде всего для исследователей. Поскольку мы рассматриваем репозиторий как часть цифровой инфраструктуры для количественных и корпусных исследований, любые публикуемые данные должны быть структурированными и машиночитаемыми. Иными словами, представленные в данных наблюдения, будь то тексты литературных произведений, сведения об изданиях или биографические сведения, должны быть в достаточной степени формализованы и способы их представления гармонизированы (то есть однотипные значения должны быть записаны единообразно), чтобы данные были пригодны для решения исследовательских задач[13]. Машиночитаемость требует также соблюдения стандартов представления данных, обеспечивающих максимальную совместимость. Мы следим, чтобы текстовые данные были представлены в кодировках Unicode, чтобы для табличных, текстовых и сетевых данных использовались стандартные форматы, подходящие для долгосрочной архивации данных и удобные для машинной обработки и обмена между системами (CSV, TXT, JSON).

Второе ключевое условие для того, чтобы данными было возможно воспользоваться, — документация. Каждый публикуемый в репозитории датасет обязательно сопровождается файлом readme, задача которого — дать представление о научной репрезентативности данных, с одной стороны, и о том, каким образом ими можно пользоваться, — с другой. Типичный файл readme должен включать сведения об источниках данных, методологии и принципах их отбора, формате и структуре входящих в датасет файлов. То есть наряду с машиночитаемостью данные должны выдерживать и достаточный уровень *человекочитаемости*.

Все датасеты до публикации проходят процедуру рецензирования. Важно отметить, что основной задачей рецензирования является корректировка, а не отсев материалов. Это не означает, что в репозитории нет отбора материалов, — он происходит на этапе рассмотрения редакцией заявки на публикацию данных. Если предложенные к публикации материалы соответствуют тематике репозитория и нашим представлениям о полезности и применимости данных, они принимаются к публикации и передаются на рецензирование.

В процессе работы мы поняли, что полноценная гармонизация данных требует формализованного тестирования данных на консистентность. В результате обязательным требованием к публикации стало сопровождение данных формальными тестами. Вот несколько типичных проблем, на отслеживание которых направлено автоматизированное тестирование: проверка на связь файлов с записями в таблице метаданных (каждому файлу соответствует запись и наоборот), проверка на отсутствующие значения в метаданных, проверка на тип значения (в качестве даты указан год, а не диапазон или знак вопроса), проверка на значения из закрытого списка, проверка на соответствие дат указанному диапазону и т.п. Обычно тесты пишутся на языке Python. Независимо от рецензирования и несмотря на большую авторскую и кураторскую работу с данными, благодаря тестированию удалось найти и устранить массу ошибок.

Таким образом, процедура публикации данных включает следующие этапы:

1. Заявка автора и решение редакции о принятии данных к публикации.
2. Оформление черновика датасета (автором при помощи редакции). Черновик доступен для просмотра только авторам, редакции и рецензентам.

---

13 См.: *Володин А.Ю.* Указ. соч. С. 21—38.

3. Рецензирование датасета, правка данных и сопроводительной документации по замечаниям рецензентов.
4. Корректура данных, подготовка автоматизированных тестов целостности данных и правка обнаруженных с их помощью ошибок и неточностей (куратор данных с сотрудничестве с автором).
5. Публикация датасета.

Подробно процедура публикации и требования к данным описаны на сайте репозитория[14].

При публикации обновления датасет уже не проходит процедуру повторного рецензирования, но обязательно проходит корректуру данных.

В отношении условий распространения публикуемых данных мы придерживаемся политики максимальной открытости. По умолчанию данным присваивается наиболее открытый тип лицензии (Creative commons — attribution, CC BY 4.0). Эта лицензия обязывает пользователей корректно ссылаться на данные, но не ограничивает их в праве пользоваться данными по своему усмотрению.

Опыт и рефлексия

Когда и как набор файлов превращается в датасет?

Такой вопрос возникает всякий раз при рассмотрении заявки на публикацию данных в репозитории. Ответ на этот вопрос требует анализа самой концепции датасета как отдельного продукта, имеющего потенциал к публикации. У нас нет простого формализованного ответа, и вряд ли он возможен. Мы руководствуемся эвристическими критериями: либо данные уже были использованы для исследования, либо коллеги (рецензенты и редакторы репозитория) видят в них потенциал для дальнейшего использования в исследовательской работе.

Однако в любом случае публикация датасета не сводится к выкладыванию набора файлов, даже если речь идет о данных к уже завершенному и опубликованному исследованию. Работа с данными — это большое поле культурных практик, которые не были в фокусе внимания исследователей в области гуманитарных наук на протяжении предшествующих десятилетий. Преобразования, которым подвергается датасет на пути от подачи заявки до публикации, во многом ориентированны на опыт, наработанный в компьютерной инженерии и точных науках. Исследователям, работающим с цифровыми материалами, нередко не хватает навыков в сфере организации данных: где файлы должны храниться и по какой системе их именовать? в каком случае сделать один файл, а в каком разбить его на множество мелких? какой формат организации данных следует использовать: базу данных, таблицу, текстовый файл? как интегрировать новые данные, если работа продолжается? почему нельзя просто сделать таблицу в Word? От принятых по ним решений зачастую зависят дальнейший жизненный цикл данных, возможность и удобство их использования и пополнения.

Обязательной частью подготовки публикации, превращающей набор файлов в датасет, является подготовка сопроводительной документации к данным — файла readme. Показательно, что мы еще ни разу не сталкивались с тем, чтобы автор данных создавал подобный файл документации для себя, безотносительно к задаче

14 https://dataverse.pushdom.ru/site/for-authors.html

публикации датасета. От краткой пояснительной записки, сопровождавшей первые публикации, мы постепенно перешли к достаточно развернутому документу, отчасти напоминающему по жанру data paper (статью, задача которой — представить коллегам опубликованный набор данных). Важное отличие наших файлов readme от data papers в том, что мы используем другую стратегию легализации данных в научно-административном поле. Data paper можно воспринимать как технологию, которая позволяет данным «притвориться» обычной статьей, чтобы их можно было цитировать и учитывать в списке публикаций, своего рода академическая мимикрия. В нашем случае данные могут просто быть собой. Тем не менее развернутая документация остается совершенно необходимой частью данных, без которых идея публикации данных для того, чтобы сделать их доступными, очень легко превращается в формальность.

Когда публиковать данные и зачем это делать?

В этом вопросе кроется главная точка расхождения утопического идеала открытой науки с реальностью. К тому, чтобы исследователи начали массово публиковать свои данные, есть препятствия разного рода. Частью они когнитивные, частью культурные, частью структурные. Эта проблема не нова и не уникальна для филологии[15]. Мотивация исследователя, который создает данные и делится ими, оказывается здесь ключевой. В миссии нашего репозитория, как она сформулирована, все заявленные цели надындивидуальны. Исследователь может десятилетиями собирать данные и тратить огромные ресурсы на их сбор. Мы же предлагаем выложить эти данные на всеобщее обозрение на благо науки в целом. Это предложение не слишком привлекательно, если исследователю кажется, что он еще не полностью исчерпал потенциал своих данных или что другие смогут с легкостью получить на основании его данных ценные результаты. Сам автор в этом случае получит лишь минимальное признание, да и то в том случае, если на его данные добросовестно сошлются. Такие опасения могут быть преувеличенными, но бывают и вполне оправданны. Редакции репозитория и всем сторонникам открытой науки необходимо поэтому искать тонкий баланс между интересами науки в целом и интересами индивидуальных исследователей.

На текущем этапе наша позиция — публиковать те данные, которые автор готов публиковать. Опыт показал, что это совсем не редкость. В определенный момент автор понимает, что сделал все, что хотел, на основании своих данных, и вполне может ими поделиться. Так в нашем репозитории появляются датасеты, которые можно назвать ретроспективными. Это материалы опубликованных работ или диссертаций, которые никогда не публиковались, хотя могли быть выложены в интернет или пересылаться коллегам. Публикацию таких датасетов мы рассматриваем как инвестицию в более долгосрочное сохранение этих данных и в большую их доступность для повторного использования. Такие датасеты, будучи опубликованными, могут быть в дальнейшем использованы научным сообществом для решения уже новых задач, отличных от тех, которые ставил автор, собравший эти данные. В настоящий момент ретроспективные публикации — важное поле для работы нашего репозитория, где мы можем внести свой вклад как минимум тем, что публикуем уже существующие в рамках дисциплины цифровые данные. Чтобы репозиторий отражал не только прошлое в работе над данными, но и настоящее, мы

15 См.: *Gomes D.G. et al.* Why don't we share data and code? Perceived barriers and benefits to public archiving practices // Proceedings of the Royal Society B. 2022. Vol. 289. No. 1987. P. 1—11. https://doi.org/10.1098/rspb.2022.1113

стремимся делать и датасеты к текущим и даже перспективным исследованиям, прежде всего за счет публикаций сотрудников Лаборатории цифровых исследований. Среди прочего, мы постановили за правило сопровождать все выполненные в лаборатории количественные исследования публикацией данных и программного кода для воспроизведения исследования. Только так, небольшими шагами, мы можем приблизиться к признанию того, что публикация данных должна стать общепринятой практикой.

Как следует делиться данными?

Оказывается, в этой области тоже есть техническое знание, заслуживающее интеграции в профессиональный инструментарий. Как было сказано выше, есть немало прецедентов, когда исследователи готовы делиться данными с коллегами. Однако мы можем констатировать, что добрая воля исследователя является необходимым, но недостаточным условием для полноценной публикации данных. Доминирующая культурная практика, которую мы наблюдаем в ситуациях передачи цифровых данных между филологами, — это хождение данных «в списках». Различные таблицы и базы данных годами пересылаются между заинтересованными исследователями в виде отдельных файлов. Если данные востребованы в течение долгого времени и в особенности если авторы продолжают свою работу над ними, естественным образом возникают различия в содержании пересылаемых файлов, стихийно складываются разные версии данных. Циркулирующие таким образом данные по существу были введены в научный оборот. Но они были введены, если воспользоваться аналогией с допечатной книжностью, в «рукописном» варианте. Если продолжить эту аналогию, публикация датасета в репозитории открывает гутенберговскую эпоху в обращении цифровых данных.

Весьма редко мы можем опубликовать в репозитории данные непосредственно в том виде, в котором получили их от автора. Иными словами, публикация датасета не сводится к выкладыванию в интернет полученных файлов. Данные всегда проходят редактирование, для того чтобы ими можно было адекватно воспользоваться. Чаще всего требуется унификация (гармонизация) данных и преобразование в более строго формализованный (машиночитаемый) формат. В ходе редакционной подготовки данные обогащаются и уточняются, тем самым становятся доступнее для дальнейшего использования.

Где публиковать данные?

У авторов есть выбор. Более молодые коллеги зачастую хорошо ориентируются в системах контроля версий, пользуются гитхабом (https://github.com), крупнейшей платформой для публикации открытого программного кода, и публикуют данные и код к своим исследовательским проектам там. Таким образом, предлагая возможность публикации данных, нашему репозиторию приходится конкурировать не только с другими публичными репозиториями научных данных, но и с гитхабом. В настоящее время наше главное преимущество, пожалуй, в редакционной подготовке — в том времени и усилиях, которые редакция репозитория вкладывает в проверку и улучшение данных до публикации. Однако приходится признать, что во многих других отношениях (вычислительные мощности, объем пространства для хранения, инфраструктура долгосрочной архивации) мы гораздо сильнее ограничены в ресурсах. Во многом стратегии выбора площадки для публикации данных остаются открытым вопросом для обсуждения в профессиональном сообществе.

Для кого публиковать данные?

Привычное представление о публикации базы данных на веб-платформе не соответствует тому, что видят пользователи в репозитории. Репозиторий дает лишь возможность скачать файлы данных. Предполагается, что, скачав данные, пользователь будет работать с ними так, как считает нужным. Это решение ориентировано в первую очередь на тех исследователей, которые работают с данными как с источником для количественных или корпусных исследований. Однако данные могут быть полезны и исследователям, использующим более традиционные филологические методы. Такие пользователи ищут на сайте репозитория возможности для просмотра данных и интерактивного взаимодействия с ними и не находят их. К сожалению, создание таких интерфейсов требует больших ресурсов, далеко выходящих за рамки наших возможностей.

В свете тех практических проблем и обстоятельств, с которыми мы столкнулись в процессе работы репозитория, следует подчеркнуть, что одним из важнейших аспектов нашей миссии оказалось формирование профессиональных практик работы с данными в ходе нашего редакционного взаимодействия с коллегами-филологами, и усилия, прикладываемые к потенциальной институализации этих практик. Кроме того, можно со всей уверенностью констатировать, что утопический образ мира открытой науки, в которой ученые сами несут свои данные для публикации в репозитории, в реалиях нашей дисциплины пока нереализуем. Исходя из этого, редакция репозитория сама ищет данные и потенциальных авторов. Если у читателей этой статьи возникла идея, что какие-то из рабочих файлов, возможно, представляют собой датасет, пожалуйста, напишите в редакцию по адресу: dataverse@pushdom.ru или напрямую главному редактору репозитория: kmaslinsky@pushdom.ru.