Месяцеслов с росписью чиновных особ в государстве на лето от Рождества Христова 1783. СПб.: Имп. акад. наук, 1783.

Российский государственный военно-исторический архив (РГВИА). Ф. 489. Формулярные списки и другие материалы о службе личного состава русской армии.

РГВИА. Ф. 490. Коллекция офицерских сказок.

Список Воинскому департаменту и находящимся в штате при войске, в полках, гвардии, в артиллерии и при других должностях генералитету и штаб-офицерам на 1784 г. СПб. : Гос. Воен. колл., 1784.

Татарников К. В. Предисловие // Послужные и смотровые списки русской армии 1730–1796 гг. в собрании РГВИА. Межфондовый указатель: В 3 т. Т. 1. / Сост., вступ. статья, оформл. К. В. Татарников. М. : Старая Басманная, 2013. С. 3–42.

УДК 004.514

А. А. Ефимов

ИНТЕРФЕЙС ДЛЯ ДАТАСЕТА[1]

В статье автор анализирует варианты публикации больших массивов данных. Приводит примеры публикации датасетов с интерфейсом для организации выборок, поиска и пр. Описывает опыт публикации датасетов разного объема и содержания, созданных в лаборатории «Международный центр демографических исследований», а также опыт коллег.

Ключевые слова: Open Access, Open Data, Dataset, Database, Interface, датасет, база данных, интерфейс.

На момент написания статьи страница, посвященная датасету (*dataset*) в русскоязычной Википедии отсутствует, но мы, не ставя перед собой задачу дать исчерпывающее определение данного термина, будем понимать под датасетом набор информации, обработанный и структурированный, представленный в табличном, иерархическом или ином виде, при условии, что вид этот формализован, описан и понятен.

Определение термина «открытые данные» можно найти в русскоязычной Википедии [Открытые данные] но в рамках данной статьи под открытыми данными мы будем понимать легально размещенный в открытом доступе датасет.

Именно содержание публикуемого контента отличает *Open Access* от *Open Data*. В первом случае речь обычно идет о публикации в открытом доступе статьи, препринта, отчета и пр., во втором – сырых или первично обработанных и описанных данных, в результате работы с которыми может появиться множество статей, препринтов, отчетов и пр. О некоторых частных случаях построения датасетов и вариантах их публикации пойдет речь в этой статье.

Первый пример, который хотелось бы показать – База данных «Екатеринбургская губернская партийная организация. 1922 год». База данных создана

[1] Тема поддержана грантом РФФИ 18-09-00592 «Эволюция крестьянской семьи на Среднем Урале в XX веке: опыт реконструкции по материалам бюджетных обследований»

на материалах Всероссийской переписи членов РКП 1922–1924 гг. (ЦДООСО, оп. 2, д. 471–517). В данном случае речь идет о датасете, который состоит из плоской таблицы, содержащей более сотни столбцов и более 12 тыс. строк, а также большого массива изображений – отсканированных копий документов (около 40 тыс. файлов). В результате систематизации информации была разработана структура описания метаданных, включающая семь таблиц.

Было принято решение обеспечить следующий функционал: возможность поиска по фамилии, имени, отчеству, дате рождения с выводом информации в общей форме, а также проведение выборки по типовым запросам. Потенциальный пользователь должен иметь возможность сформировать свой запрос на выборку с учетом решаемых задач.

Вопрос публикации такого объема материалов, а также организации структуры просмотра и поиска нетривиален. Вариант публикации плоской таблицы и архива с изображениями не рассматривается в силу того, что подобным датасетом будет просто невозможно пользоваться.

Нами было принято решение адаптировать *CMS DSpace* для реализации точки публикации данного датасета. Метаданные и данные были обработаны соответствующим образом, а база данных зарегистрирована [База данных «Всероссийская перепись»] и опубликована онлайн [Всероссийская перепись]. Поставленные задачи в целом реализованы. Датасет, лежащий в основе базы данных, интерфейс которой реализован на *DSpace*, доступен онлайн, а интерфейс, равно как и его кастомизации, документирован и несложен в поддержке и обновлении. Все компоненты интерфейса, такие как операционная система, сервер СУБД, Веб-сервер и пр. доступны под свободными лицензиями.

Во многом аналогичное решение использовано коллегами из Пермской государственной краевой универсальной библиотеки для публикации набора данных, посвященного династии Романовых, коллегами из «Лаборатории эдиционной археографии» в рамках создания «Электронного корпуса исторических источников»[1] и пр.

В результате создания подобных датасетов был получен опыт описания, обработки и публикации больших объемов сложного контента, опыт оптимизации и кастомизации *CMS DSpace*. Были сделаны выводы о возможности применения подобного подхода для публикации датасетов, содержащих, помимо большого объема текстовых табличных данных, еще и десятки гигабайт графических и мультимедийных материалов.

Решения, применяемые в случае объемных текстовых и мультимедийных данных неприменимы в случае, когда датасет представляет собой плоскую

[1] Электронный корпус создается в рамках проекта реализации проекта «Возвращение в Европу: российские элиты и европейские инновации, нормы и модели (XVIII – начало XXв.)», дог. № 14.A12 31.0004 от 26.06.2013 г.

таблицу, содержащую сотню строк и сотню столбцов. В случае компактного датасета, который стал результатом обработки большого объема первичных данных и работы эксперта, ту же *CMS DSpace* использовать смысла нет. Нами было принято решение обеспечить к датасету онлайн доступ и реализовать для этого доступа интерфейс. База данных «Крестьянские хозяйства Уральской области. 1928/1929» [База данных «Крестьянские хозяйства»] составлена на основе формуляра «Бланк бюджетного описания крестьянского хозяйства. 1928/1929 г.» и содержит 325 описаний [хозяйств], что соответствует примерно 1 800 строкам. Каких-то мультимедийных, графических файлов и пр. материалов нет – только текст.

Для удобства публикации и оперативного обновления было принято решение использовать табличный процессор для ведения базы данных с возможностью экспорта результатов в удобный машинно-читаемый формат (*XML*) и внешний стилевой фильтр для реализации выборки (по описаниям) и минимального визуального оформления таблицы. Обычно подобные датасеты публикуются онлайн просто как плоская таблица, без возможности встроенного поиска, сортировки и пр. Мы же реализовали не только онлайн доступ к датасету, но и интерфейс для работы с ним.

В заключение хотелось бы отметить, что тенденция перехода от *Restricted Access* к *Open Access* в научной среде есть, и практически ни у кого не вызывает вопросов. Тенденция обогащения *OpenAccess* посредством *Open Data* тоже есть, но вопрос реализации концепции *OpenData* в случае публикации нетипичных, уникальных, сложных датасетов остается открытым.

Список литературы

База данных «Всероссийская перепись членов РКП(б). 1922–1923 гг.» : свидетельство о регистрации. URL: http://elar.urfu.ru/handle/10995/82059 (дата обращения: 09.08.2020).

База данных «Крестьянские хозяйства Уральской области. 1928/1929 [Электронный ресурс] // Научная лаборатория «Международный центр демографических исследований. URL: https://idun.urfu.ru/ru/pro/ehvoljucija-krestjanskoi-semi-na-srednem-urale-v-xx-veke/baza-dannykh/baza-dannykh-krestjanskie-khozjaistva-uralskoi-oblasti-19281929/ (дата обращения: 09.08.2020).

Всероссийская перепись членов РКП (б). 1922–1923 г. [Электронный ресурс] // Научная лаборатория «Международный центр демографических исследований. URL: https://idun.urfu.ru/archive/ (дата обращения: 09.08.2020).

Открытые данные [Электронный ресурс] // Википедия. URL: https://ru.wikipedia.org/wiki/%D0%9E%D1%82%D0%BA%D1%80%D1%8B%D1%82%D1%8B%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5 (дата обращения: 09.08.2020).

Центр документации общественных организаций Свердловской области (ЦДООСО). Ф. 76. Екатеринбургский губернский комитет РКП (б).