Открытые данные

В. В. Староверов

***Аннотация* - Настоящая работа посвящена обзору концепции и структуры открытых данных, состоянию развития данного направления, имеющихся положительных сторонах и недостатках, а так же возможных решениях. Основная задача статьи - рассмотреть текущее положение дел в сфере открытых данных, имеющиеся трудности и способы их преодоления. Так же будет описана возможная модель и архитектура портала открытых данных, которая, по мнению автора, является продолжением развития данного направления.**

Ключевые слова - Открытые данные, порталы открытых данных, наборы данных, машиночитаемые данные.

I. ВВЕДЕНИЕ

Концепция открытых данных на сегодняшний день не является чем-то новым и уникальным, при том, что формализация данного определения относительно новая. Основная мысль данной концепции лежит в термине «Открытое определение», которое можно в общих словах описать следующим: «Часть информации является открытой только в том случае, если любой желающий может свободно использовать, повторно использовать и распространять данные» [2].

Цели движения открытых данных похожи на другие «открытые» движения, такие как открытое программное обеспечение (open source), открытый контент (open content) и открытый доступ (open access). Рост популярности идеи об открытых данных во второй половине 2000-х годов связан, прежде всего, с запуском правительственных инициатив, таких как Data.gov.

Открытые данные часто ассоциируются с нетекстовыми материалами, такими как карты, геномы, химические компоненты, математические и научные формулы, медицинские данные, данные о биологическом разнообразии. Проблемы чаще всего возникают по той причине, что эти данные могут быть коммерчески ценными или могут быть собраны в некие ценные продукты [1].

Доступ к данным, как и последующее их использование, контролируется организациями — государственными и частными. Контроль может быть через ограничения, лицензии, копирайт, патенты и требования оплаты для доступа или повторного использования. Сторонники идеи «открытых данных» считают, что подобные ограничения идут против общественного блага и данные должны быть доступны без ограничений или оплаты. Также важно что данные должны быть доступны без последующих запросов на разрешение, хотя и способы повторного использования, такие как создание продуктов на базе данных, могут контролироваться лицензией [2].

Государственные данные представляют один из ключевых интересов для общества, и многочисленные некоммерческие организации и отдельные активисты добиваются открытости государственной информации в машиночитаемой форме. Многие национальные правительства в рамках стратегий «открытого государства» создали веб-сайты для распространения части данных, обрабатываемых в секторе государственного управления [1].

II. ОТКРЫТЫЕ ДАННЫЕ

Термин «Открытое определение» дает полное описание требования к открытым данным и содержимому этих данных. Открытые данные являются одним из блоков так называемых открытых сведений. Сами открытые сведения являются квинтэссенцией полезности и применяемости открытых данных [4].

Ключевые особенности открытых данных следующие:

- **Доступность.** Данные должны быть доступны в целом и иметь разумную цену воспроизводства, опирающуюся на доступ к данным средствами Интернет. Данные должны быть доступны в удобной и изменяемой форме;
- **Повторное использование и распространение.** Данные должны предоставляться на условиях повторного использования и распространения, включая объединение с другими наборами данных. Данные должны быть машиночитаемыми;
- **Всеобщее участие.** Любой вправе использовать, повторно использовать, распространять открытые данные. Не должно быть никаких препятствий, мешающих выполнению этих условий для человека или группы людей. Например, "некоммерческие" ограничения, которые бы препятствовали «коммерческому» использованию, или ограничения пользования для определенных целей (например, только в образовании), не допускаются.

A. ТИПЫ ОТКРЫТЫХ ДАННЫХ

На сегодняшний день существует множество типов открытых данных, потенциально используемых. Некоторые из них:

- **Культурные,** такие как сведения о раскопках и артефактах, музеях, культурных мероприятиях и прочее;
- **Научные.** Данные, публикуемые как часть научных исследований, начиная от астрономии и заканчивая зоологией;

В. В. Староверов — магистр факультета ВМК МГУ имени М. В. Ломоносова (email:oldbelieve@mail.ru)

- **Финансовые,** такие как сведения о правительственных счетах (доходы и расходы государства) и сведения о состоянии финансовых рынков;
- **Метеорологические.** Данные, которые бы позволили следить за изменениями в погоде и климате;
- **Статистические,** т.е. данные, полученные с помощью статистических управлений, таких как переписи населения и основные социально-экономические показатели;
- **Данные об окружающем пространстве.** Информация, связанная с окружающей природной средой, например, присутствия и уровень загрязняющих веществ, качество рек и морей;
- **Данные о транспорте.** Всевозможные сведения о маршрутах, расписания движений и т.д.;

B. НЕОБХОДИМОСТЬ ОТКРЫТЫХ ДАННЫХ

Ответ на данный вопрос основывается на типе сведений в открытых данных. И все же, основные причины следующие:

- **Прозрачность.** В корректно функционирующем демократическом обществе, граждане должны знать, что делает их правительство. Для этого они должны иметь возможность свободно получать доступ к правительственным данным и информации, а так же обмениваться информацией с другими гражданами. Прозрачность должна быть не только в доступе, но так же и в совместном использовании. Часто, чтобы понять материал, его необходимо проанализировать и визуализировать, а это в свою очередь требует от материала быть используемым, повторно используемым и распространяемым;
- **Свободная социальная и коммерческая ценность.** В век цифровых технологий, данные являются ключевым пунктом социальной и коммерческой деятельности. Все, начиная от нахождения вашего местного почтового отделения, для построения поисковой системы требует доступа к данным, большая часть которых создается или предоставляется правительством. Открывая данные, правительство может помочь управлять созданием инновационного бизнеса и услуг, имеющие социальную и коммерческую ценность;
- **Участие и взаимодействие.** Большинство граждан имеют возможность взаимодействовать с правительством лишь эпизодически, а то и вовсе только на выборах, проходящих раз в 4 - 5 лет. Открывая данные, правительство предоставляет гражданам возможность быть более информированными и принимать участие в процессе принятия решений. И это даже больше чем прозрачность: это полный доступ к обществу, не только знание того, что происходит в деятельности власти, но и возможность внести свой вклад в процесс развития общества.

Таким образом, можно утверждать, что развитие концепции открытых данных, ее повсеместное внедрение является шагом на пути к процветающему самоконтролируемому демократическому обществу [4].

III. РЕШЕНИЯ ДЛЯ ОТКРЫТЫХ ДАННЫХ

Согласно Федеральному закону от 9 февраля 2009 г. № 8-ФЗ «Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления» органы местного самоуправления обязаны предоставлять всем заинтересованным сведения о своей деятельности. Одним из способов реализации федерального закона органами местного самоуправления является публикация открытых данных по тематике деятельности органа власти на своих ресурсах в сети Интернет [3].

Данные, предоставляемые ОГВ (органы государственной власти) должны быть предоставлены в виде машиночитаемых документов и быть доступными по определенным ссылкам.

У этого подхода имеется один главный недостаток: пользователям необходимо самостоятельно искать нужные данные, при этом, какого-либо универсального подхода в публикации этих данных нет.

Для решения описанной выше проблемы, а так же ряда других, таких как актуализация данных, единый реестр и интерфейс и прочее, применяют порталы открытых данных.

Портал открытых данных (ПОД) - Web-ресурс в задачи которого входит хранение, версионирования, катологизация, актуализация, сбор и предоставление доступа к открытым данным.

Общая архитектура почти каждого портала открытых данных представлена на рисунке 1.

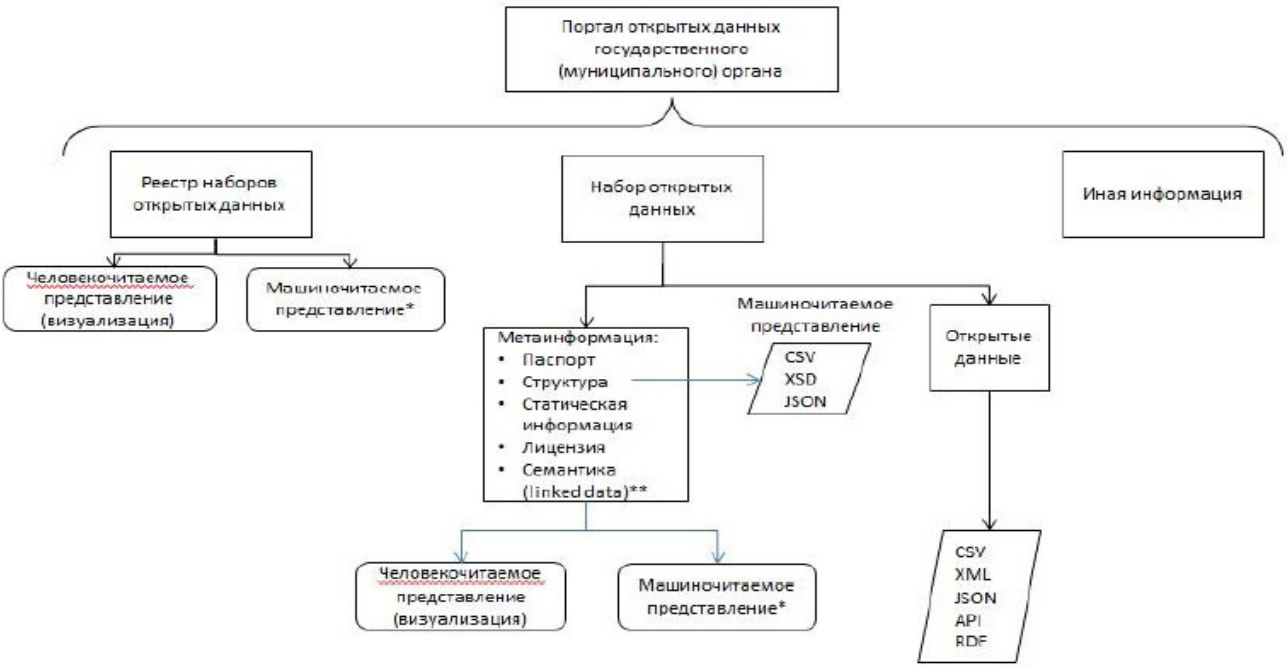

Рис. 1 Архитектура портала открытых данных

При разработке данной статьи были изучены ресурсы порталов открытых данных из следующего списка (первые три из выдачи поисковой машины Google по запросу "Портал открытых данных"):

1) Федеральный портал открытых данных [5];
2) Портал открытых данных г. Москва [6];
3) Портал открытых данных «Открытые данные Краснодар» [7].

Назначение разрабатываемого ресурса определяется как региональное решение для выполнения 4, 7, 9, 13 и 14 статей Федерального закона "Об обеспечении доступа к информации о деятельности государственных органов и органов местного самоуправления". Данное постановление вводит в оборот так же обязательные требования к размещению информации на порталах открытых данных, доступ к ним и другое, называемое методическими рекомендациями. Разрабатываемое решение полностью соответствует данным методическим рекомендациям, а также расширяет и дополняет их.

Разрабатываемый ПОД представляет собой ресурс в сети интернет, предоставляющий пользователям доступ к открытым данным через следующие интерфейсы:

- Графический интерфейс, реализованный средствами HTTP и доступный через Web браузер;
- Предоставлением HTTP REST API для мобильных приложений;
- Предоставление доступа к файлам наборов средствами HTTP GET запросов и FTP.

При разработке архитектуры приложения основной упор шел на простоту использования как для поставщиков информации, так и для ее конечных потребителей, выделяя таким образом данное решение среди остальных, о чем будет сказано далее.

Основное отличие разрабатываемого портала открытых данных для конечного потребителя от остальных решений - это быстрый доступ к необходимой информации, строго заданная ее структура и содержание. Для примера, если рассмотреть федеральный ПОД и портал города Москва, можно заметить, что все наборы на нем представлены в виде обычных таблиц, информация в которых отображается не корректно в виду того, что данные отображения не имеют структуры, а сам портал не предоставляет об этой структуре никакой информации. Аналогичное происходит и при доступе к API данных ресурсов. Таким образом, разработчиком приложений, который предполагают интеграцию своих программных продуктов с данными ресурсами, необходимо знать назначение наборов данных, их структуру и состав для адекватного предоставления информации пользователям. При этом, в методических рекомендациях предполагается версионирование наборов данных с расчетом не только на изменение состава набора, но и на полное переопределение его структуры в следствии чего, разработчикам приложений приходится выпускать обновления своих программных продуктов чтобы учитывать все внесенные изменения. При этом, некоторые наборы могут обновляться настолько часто, что разработчики просто не в состоянии будут сопровождать свой продукт, а пользователям быстро надоест регулярное обновление.

Таким образом, рассматриваемые решения не доступны для наборов с частой сменой структуры. Разрабатываемое решение в первую очередь опирается на то, что у набора данных есть структура, и эта структура строго определена и детально описана самим поставщиком. Создание описания просто для поставщика, так как он использует графические конструкторы, предоставляемые порталом, и с другой стороны имеет достаточно информации для того, чтобы приложение само могло его использовать при работе с набором данных.

Версионирование наборов данных так же является отдельным аспектом. Методические

рекомендации определяют тот факт, что набор может меняться с течением времени, а пользователям может потребоваться доступ к более старой информации. Рассматриваемые решения поддерживают версионирование наборов. Разрабатываемое решение так же поддерживает данный функционал, однако расширяет его возможностью древовидной организации версий набора, когда из одного набора может быть выдвинуто две параллельные версии. Данный функционал может показаться ненужным в виду того, что поставщик у набора один, и информация у него так же единственная, однако, в случае, если набор содержит информацию о возможных прогнозах и их исходах, опираясь на данные и предыдущей версии, такое представление версионирования набора вполне оправдано.

Отдельно стоит сказать о структуре данных. Некоторые наборы данных предполагают наличие связанных в отношении 1:n данных. К таким наборам данных относятся предоставляемые ПОД «Открытые данные Краснодар».

Очевидно, что данная информация может быть представлена в виде древовидной структуры. Портал открытых данных Краснодара поддерживает данные структуры в отличии от остальных рассматриваемых, и предоставляет ее пользователю в красиво структурированном виде. Остается неизвестным, как происходит внесение таких наборов на портал и как он с ними работает. Предлагаемое решение за счет выделение структуры данных позволяет точно определять структуру, состав и типы данных полей наборов.

Как видно из архитектуры, портал предоставляет пользователям доступ к открытым данным при помощи одного из двух интерфейсов:

1) **Человекочитаемое представление.** Это интерфейс, позволяющий пользователю просматривать набор открытых данных прямо на портале в удобной для понимания форме;
2) **Машиночитаемое представление.** Это электронный документ в одном из форматов, применяемых для обмена данными внутри сетей между машинами с разными архитектурами. В большинстве своем, это форматы основанные на текстовых файлах, например, CSV или XML. Данный интерфейс делится на 2 типа:

a) **Файловый.** Доступ непосредственно к машиночитаемому файлу через определенную ссылку;
b) **REST API.** Доступ к данным средствами Web-API, предоставляемого порталом.

Оба интерфейса используются для взаимодействия сторонних устройств и систем с порталом открытых данных.

Есть так же важные элементы этой архитектуры. Остановимся на основных из них более подробно:

- **Реестр наборов открытых данных.** Реестр представляет собой хранилище открытых данных, где сами наборы представлены в определенной иерархии, имеют версионирование. Назначение реестра - так организовать хранение данных, чтобы доступ к нему был максимально прост и эффективен по любому из предоставляемых порталом интерфейсов.
- **Паспорт.** Паспорт - это машиночитаемый документ, в назначение которого входит явная идентификация набора данных в реестре. Паспорт содержит в себе сведения о наборе данных, информацию о публикаторе набора, ссылки на набор и его версии, а так же сведения о владельце и периодах актуализации данных в наборе. В основном, поиск набора данных любой внешней системой, или человеком, начинается именно с поиска паспорта набора. К паспорту набора предъявляется определенная структура, и он является машиночитаемым документом.
- **Структура.** В большинстве случаев, это документ с метаинформацией по набору, описывающей то, как этот набор отображать пользователю, либо какие-то данные, требуемые машинам для работы конкретно с данными набора.
- **Статистическая информация.** Это информация о количествах скачивания, использования, просмотра набора. Так же содержит в себе информацию, оставляемую пользователем, такую как оценки и достоверность. В большинстве случаев применяется для оценки востребованности набора и адекватности содержащейся в нем информации.

На сегодняшний день, все наборы данных предоставляются пользователям в виде простых таблиц содержащих текстовую информацию. Внешний вид наборов остается примитивным и зачастую не соответствует их реальному виду. Среди рассмотренных ПОД так же не обнаружилось какой-либо сущности, которая бы описывала структуры наборов. Среди рассмотренных решений все данные были представлены в виде CSV страниц. Данный нюанс крайне негативно влияет на усваиваемость информации пользователем, а возможность использования данных наборов сторонними системами затруднительна из-за того, что разработчики данных систем могут не знать о назначении отдельных элементов данных.

Ко всему прочему, стоит упомянуть о самой информации размещаемой на порталах. В первую очередь, вся информация должна быть машиночитаемой. Таким образом все типы данных предоставляемой информации являются простыми, такими как текст, числа, URL, и прочие, которые явно могут быть представлены в виде текста. Большинство ПОД расширяют этот перечень тем, что предоставляет карту с геометками, в случае, если набор содержит геолокационную информацию. Определение «машиночитаемый» подразумевает тот факт, что вычислительная машина может обрабатывать эти данные. Современный уровень развития вычислительной техники и программного обеспечения достаточно высок, чтобы расширить список машиночитаемых форматов до мультимедийной информации, такой как изображения, аудио- и видеофайлы.

IV. ПРЕДЛОЖЕНИЯ ПО РАЗВИТИЮ АРХИТЕКТУРЫ ПОДТАЛА ОТКРЫТЫХ ДАННЫХ

Данный раздел статьи рассматривает аспекты позволяющие улучшить и расширить текущую архитектуру порталов открытых данных.

Как было описано в предыдущем разделе, у ПОД отсутствует какое-либо описание структуры данных, хранящихся в наборе. Это означает, что данные не имеют какого-то определенного формата. Это, пожалуй, наиболее значительное упущение при разработке архитектуры. В качестве предлагаемого решения данной задачи предлагается определить дополнительную сущность, а именно структуру данных.

В задачи структуры данных должно входить не только описание полей набора, но и их тип данных и их иерархию внутри набора. Это позволит в автоматизированном режиме формировать корректное отображение набора данных для пользователя, а так же предоставит функционал, за счет которого можно будет публиковать наборы данных, структура которых не может быть представлена в виде простых таблиц, либо такое представление затруднит восприятие набора данных. Примерами таких наборов могут служить сведения о структурах ОГВ, которые являются обязательными к публикации. В виду того, что данные наборы могут содержать сущности с отношениями многие ко многим, представление их в виде таблицы уже становится невозможным. Определяя четкую структуру данных, мы можем реализовать возможность ссылаться с одной сущности на другую, тем самым, реализуя отображение данных в виде взаимосвязанных таблиц. Так же четкое определение структуры данных позволит строить отображения наборов в виде сложных таблиц, содержащих вложенные ячейки и таблицы тем самым, делая работу с набором более очевидной для конечного потребителя.

Дополнительным аспектом служит расширение уже имеющихся структур данных, добавляя в них новые типы, такие как мультимедийная информация. Это позволило бы более детально формировать содержимое набора.

Отдельного обсуждения так же требует процесс публикации новых и актуализации уже имеющихся наборов данных. Оба процесса требуют наличия двух участников - оператор системы и поставщик данных. Оператор - это человек имеющий доступ к определенной административной части портала, в задачи которого входит группирование наборов и создание структур данных для наборов. Поставщик данных - это лицо или электронный сервис, в задачи которого входит предоставление данных для наборов открытых данных. В случае если поставщиком данных является электронный сервис, разрабатываемое решение автоматически инициализирует поставку данных путем взаимодействия с системой-поставщиком средствами Microsoft SSIS. В остальных случаях поставщик предоставляет данные системе путем передачи ей файлов со структурой, определенной оператором и утвержденной поставщиком.

Процесс публикации нового набора выглядит следующим образом. Поставщик отправляет заявку оператору портала ОД на публикацию набора, в которой обосновывает необходимость размещения набора на портале, содержание набора и его примерную структуру. Оператор выделяет на портале место под набор в определенной категории данных, создает структуру данных и согласовывает ее с поставщиком. В случае успешного согласования оператор предоставляет доступ поставщику к набору данных. Поставщик в своей административной части портала может получить доступ к набору для внесения данных. Внесение данных может быть осуществлено через графический интерфейс портала, либо путем отправки на портал сгенерированного по структуре набора Excel документа, который заполняет пользователь. После внесения данных, набор помечается как черновик и ожидает валидации оператором и ответственным лицом от поставщика. Если набор проходит валидацию, оператор размещает его на портале, в противном случае отправляет обратно на доработку поставщику.

Если в роли поставщика выступает другая информационная система, оператор так же создает структуру данных и сервис Microsoft SSIS, задачи которого - преобразование данных системы-поставщика в структуру, определенную оператором. При этом, поскольку данные в системе-поставщике уже проверены и утверждены, преобразованная SSIS информация автоматически публикуется на портале.

Обновление набора - процесс аналогичный публикации, за тем исключением, что не создается структура, и оператор не участвует в процессе.

Таким образом, на портале организовано четкое взаимодействие между поставщиками данных и наборами. Портал автоматически отслеживает состояние наборов и оповещает поставщиков о необходимости актуализации данных, опираясь на данные из паспорта набора.

V. ЗАКЛЮЧЕНИЕ

Открытые данные представляют один из ключевых интересов для общества и многочисленные некоммерческие организации и отдельные активисты добиваются открытости государственной информации в машиночитаемой форме. Многие национальные правительства в рамках стратегий «открытого государства» создали веб-сайты для распространения части данных, обрабатываемых в секторе государственного управления.

Поэтому так важно развивать это движение. Развитие открытых данных позволит гражданам получать актуальную и достоверную информацию об интересующих их аспектах жизни государства и работы его правительства. Открытые данные позволят гражданам активно принимать участие в жизни страны, определяя курс ее развития в направлении, удовлетворяющем требования большинства граждан. Порталы открытых данных как раз являются унифицированным инструментов в достижении этих целей. Именно этот фактор обуславливает необходимость развития технологий ПОД, делая их, по возможности, наиболее универсальными.

Рассмотренные идеи по модернизации ПОД как раз направлены на максимальную универсализацию и автоматизацию процессов жизнедеятельности открытых данных. Таким образом, увеличивая актуальность, достоверность и доступ к открытым данным.

БИБЛИОГРАФИЯ

[1] Википедия: Открытые данные, https://ru.wikipedia.org/wiki/Откртытые_данные
[2] The Open Defenition, http://opendefinition.org/
[3] Методические рекомендации по публикации открытых дан- ных, 3 изд. Data.gov.ru, 29 мая 2014 г.
[4] Auer, S. R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. (2007). "DBpedia: A Nucleus for a Web of Open Data". The Semantic Web. Lecture Notes in Computer Science 4825
[5] Data.gov.ru. Открытые данные России, http://data.gov.ru/
[6] Портал открытых данных Правительства Москвы, http://data.mos.ru/
[7] Портал открытых данных «Открытый Краснодар», http://opendata.krd.ru/

On Open Data

V.V. Staroverov

Abstract — **This work is devoted to reviewing the concept and structure of open data, the state of development of this direction, the existing positive aspects and disadvantages as well as possible solutions. The main objective of this article is to discuss the current state of Affairs in the field of open data, existing difficulties and ways of overcoming them. Also, it describes the possible model and the architecture of the open data portal, which, according to the author, is a continuation of development in this direction.**

Keywords — Open Data, Open Data Web Applications, Data Sets, machine-readable data.