05.09.2013  Лекция №1

**Открытые данные: введение**

**Часть 1**

**Иван Бегтин**

Я вначале скажу краткое вступительное слово о том, кто мы, что мы, как это началось, и о чем будет сегодняшняя лекция. Мы, то есть некоммерческое партнерство «Информационная культура», при поддержке Открытого Правительства и, отчасти по его инициативе, мы открываем «Школу открытых данных» в России. Это просветительский и образовательный проект, в котором мы будем рассказывать о том, что такое открытые данные и как с ними работать. У нас на занятиях будет как базовый теоретический материал, так и материал очень практический, т. е. мы будем показывать, как именно собирать данные, как их визуализировать, как с ними работать далее. У нас будет множество лекций, а также какое-то количество мастер-классов. Часть наших лекций мы хотим сделать публичными. У нас заявок на участие было, честно говоря, сильно больше, чем мы смогли позвать сюда людей, потому что зал здесь у нас не очень большой. Надеюсь, для последующих лекций нам удастся найти зал побольше. А некоторые лекции мы хотим вообще сделать публичными, чтобы на них могли прийти вообще все желающие.

Итак, сегодня будет вступительная лекция, состоящая из двух частей. Первую прочитает Ирина Радченко, доцент и сотрудник Высшей школы экономики. Ирина - специалист по открытой науке, открытому доступу и открытым данным применительно к технологиям связанных данных. Это очень интересная тема, которую мы будем в перспективе обсуждать более углубленно. В первой части будет вводный обзор того, что такое вообще открытые данные.

Во второй части я буду рассказывать об открытых данных в международной практике том, как люди во всем мире работают с открытыми данными, о том, где почерпнуть информацию, где узнать больше, в каких сообществах имеет смысл поучаствовать, откуда получать самую последнюю актуальную информацию и где находятся обучающие материалы, которые все мы используем. В конце своей лекции я проведу небольшой мастер-класс касательно Хаба открытых данных. На примере одного конкретного портала открытых данных, развернутого нами, я покажу, как с ним работать на практике. После этого вам будет дана часть задания, о котором я уже подробнее расскажу во время своей лекции.

Теперь передаю слово Ирине.

**Ирина Радченко**

Иван меня уже представил, так что я не буду повторяться и перейду сразу к интересному, а самое интересное, о чем я хочу вам рассказать, это открытые данные.

Это то, что входит в сферу моих профессиональных интересов, и то, чем я занималась на протяжении последнего года очень интенсивно. Прежде всего, хотела бы акцентировать внимание на том, что лекция у меня вводная и обзорная, соответственно о многих вещах я буду говорить по верхам. Но, тем не менее, в моей презентации будут ссылки на тематические информационные источники и на образовательные ресурсы. Ссылки на дополнительную литературу я уже выложила в блоге «Школы открытых данных». <http://opendataschool.ru/2013/09/introduction-to-open-data/>

Итак, я хотела бы начать с краткого обзора мирового движения открытых данных. Надо понимать, что открытые данные зародились не в России, т. е. это мировое движение. Открытые данные — это фактически новая информационная реальность. В 2009 году ИТ-директор

администрации президента США Вивек Кундра запустил федеральный портал открытых данных. Вслед за этим был запущен федеральный портал открытых данных в Великобритании. По всему миру стали запускаться порталы открытых данных, и там стали публиковаться наборы данных. На данный момент времени некоторое количество порталов открытых данных представлены, как в Европе, так и в Австралии, и в Африке, и в Южной Америке. Есть и в России, но пока что не федерального значения.

**Что такое открытые данные?**

Итак, что же такое открытые данные? Это данные, представленные в машиночитаемом виде и доступные для повторного использования без ограничений со стороны публикатора. Эти данные выложены в Интернете под открытыми лицензиями, которые позволяют их повторно использовать. Есть несколько видов открытых лицензий, в том числе лицензии, разработанные различными международными организациями: Creative Commons (<http://creativecommons.org/choose/>),  Open Data  Commons (<http://opendatacommons.org/licenses/>), но есть также открытые лицензии, разработанные государственными органами, как например Open Government Licence ([http://www.nationalarchives.gov.uk/doc/open-government- licence/version/1/open-government-licence.htm](http://www.nationalarchives.gov.uk/doc/open-government-)), разработанная правительством Великобритании.

Открытые данные обязательно должны быть представлены в машиночитаемых форматах. Это может быть формат XLS или CSV, KML, XML, JSON, RDF или другой машиночитаемый формат. Надо понимать, что все эти форматы разные и они предназначены для разных целей.

**Пятизвездочная модель Тима Бернерса-Ли**

В 2010 году Тим Бернерс-Ли разработал пятизвездочную модель открытых данных (<http://www.w3.org/DesignIssues/LinkedData.html>). Он условно разделил данные на 5 категорий открытости (<http://5stardata.info/>). Первой ступени открытости данных он присвоил 1 звездочку. Это просто данные, предоставляемые в интернете под открытыми лицензиями (например, в формате PDF). Две звездочки - это уже когда данные представлены в машиночитаемых форматах и пригодны для повторного использования, что очень важно, и к этому мы обязательно вернемся. Следующий этап (три звезды) - это когда данные представлены в непроприетарных, то есть в открытых форматах. Если на предыдущей ступени вполне годилось использование формата Excel для таблиц, то на этом этапе мы уже переходим к формату CSV. На четвертой ступени (4 звезды) формат данных подразумевает, что данные выкладываются в RDF, стандартизацией которого занимается консорциум W3C. И 5 уровень - это когда предыдущие данные, выложенные в формате по модели представления данных RDF, объединяются в так называемые пространства связанных открытых данных (Linked Open Data). То есть, по сути, мы перед собой видим огромнейшую распределенную базу данных из разнородных информационных источников.

И с этого ракурса мы можем проследить эволюционирование веба. Сначала были представлены документы в вебе, затем - представление веба в виде гипертекста и гипертекстовых ссылок. И следующий этап - веб данных (Web of Data) - это когда открываются данные и затем эти данные объединяются в единое международное виртуальное пространство связанных данных (Linked Data).

Надо сказать, что тенденция к увеличению этого пространства нарастает весьма большими темпами. На данный момент связанные открытые данные  охватили самые различные области знаний.

**Публикаторы и пользователи**

Теперь, если мы вернемся к открытым данным и посмотрим на них с точки зрения того, а что же это такое, что же за информационная среда, в которой они находятся, что они позволяют делать, какие новые возможности они предоставляют людям, - то мы можем рассмотреть всю эту среду с точки зрения двух сторон. С одной стороны находятся публикаторы данных. Это могут быть органы исполнительной власти, различные научные организации, граждане и так далее. С другой стороны, это пользователи, которые применяют открытые данные в своей работе и тем самым создают обратную связь по использованию открытых данных. При грамотном применении этой обратной связи можно существенно повысить ценность этих данных. То есть разработчики программного обеспечения, обрабатывая сырые данные, могут отследить ошибки, возникающие при их использовании, сообщить об этом публикатору, а публикатор, в свою очередь, сможет скорректировать неточности этих данных.

**Экосистема открытых данных**

Это все является экосистемой открытых данных. Очень важно понимать, что открытые данные это не явление, находящееся где-то в вакууме, и искусственным образом созданное. Это все является единой информационной средой, в которой есть разработчики и пользователи открытых данных, есть люди, которые занимаются обучением, обязательно присутствуют представители бизнеса, которые тоже используют эти данные в своих целях. Вся эта экосистема должна работать таким образом, чтобы совершенствоваться и предоставлять все больше возможностей по использованию открытых данных.

Таким образом, у нас получается следующая цепочка: обучение рядовых пользователей (хотя бы даже и “с нуля”) или, может быть, более продвинутых пользователей (например, программистов), но именно с акцентом на использование и предоставление данных; затем - использование данных в различных целях; и, наконец, это все объединяется в единую экосистему.

**Источники открытых данных**

Где же искать открытые данные в интернете? Да, открытые данные уже сейчас представлены в Сети. Я условно поделила представленные в интернете открытые данные на 3 типа. Прежде всего, это порталы международных организаций, собирающие у себя данные, поступающие, в том числе, и от официальных государственных органов; далее, это государственные порталы федерального уровня; а также порталы, созданные усилиями гражданских активистов. Надо понимать, что эта классификация весьма условная, но в общих чертах она описывает ситуацию в применении к любой стране, где есть движение открытых данных.

**Примеры порталов**

Я приведу несколько примеров, для того чтобы вы могли дома проработать материал, посмотреть порталы, в том числе, и международных организаций. Итак, это крупная международная Организация Экономического Сотрудничества и Развития (<http://www.oecd.org/statistics/>), это также ООН — Организация Объединенных Наций (<http://data.un.org/Default.aspx>), это Всемирный Банк (<http://data.worldbank.org/>) и другие организации.

Есть замечательный ресурс <http://datos.fundacionctic.org/sandbox/catalog/faceted/>, содержащий в себе список каталогов порталов открытых данных. Здесь можно посмотреть порталы открытых данных со всего мира. Кстати, обратите внимание, что здесь применяется классификация открытости “по звездам”, т. е. можно посмотреть, как данные представлены в формате Linked Open Data (связанных открытых данных), так и более низкие уровни публикации;  и вообще посмотреть, как устроены эти порталы, в каком виде выложены дата-сеты (наборы данных), поизучать конкретные каталоги открытых данных. Например, каталог открытых данных Евросоюза (<http://open-data.europa.eu/>); или федеральные государственные порталы - например, каталог открытых данных США (<https://explore.data.gov/catalog/raw>), каталог государственных открытых данных Великобритании (<http://www.data.gov.uk/data>), каталог открытых данных Чехии (<http://www.opendata.cz/>) и другие каталоги.

И, наконец, мы пришли к порталам открытых данных, которые выкладывают гражданские активисты. У нас в России такой портал открытых данных присутствует, это Хаб открытых данных <http://hubofdata.ru/>, о котором Иван расскажет в подробностях чуть позже.

**Представление данных**

Надо сказать, что все эти порталы открытых данных стремятся работать в соответствии с некоторыми... ну, пока это еще сложно назвать стандартами, но, тем не менее, экосистема открытых данных постоянно совершенствуется в направлении некой унифицированности. В частности, это происходит на уровне разработки все более усовершенствованного программного обеспечения. Большинство государственных порталов федерального уровня представлены в настоящий момент системой управления данными CKAN (<http://ckan.org/>), разработанной в организации Open Knowledge Foundation (<http://okfn.org/>). На эту систему перешли, в числе прочих, и порталы открытых данных США и Великобритании, два флагмана в  движении открытых данных. Существуют также очень известные мультимодельные гибридные системы вроде Virtuoso (<http://virtuoso.openlinksw.com/>). Эта система специально предназначена для работы именно со связанными открытыми данными. Если система CKAN ориентирована просто на представление данных, то Virtuoso — это уже следующий шаг по представлению  связанных данных.

**Процесс работы с данными**

Весь процесс работы с открытыми данными я условно поделила на 4 части. Это, прежде всего, получение данных - то, о чем я уже рассказала. Следующие ступени — это обработка и анализ данных. Это то, что нам предстоит изучить глубже в последующих лекциях. И это уже как раз та часть, за которую ответственны сами пользователи открытых данных, то есть все мы, граждане. И, наконец, представление результатов работы с данными. Это самое интересное: что же мы можем получить из открытых данных? Какую историю можем рассказать на основе открытых данных? Как ее представить? Будет ли это текстовый формат? Или это будет история в визуальном формате? В любом случае это будет аналитический продукт.

**Сферы применения открытых данных**

Я хочу подчеркнуть, что существует множество вариантов использования открытых данных. Прежде всего, это применение в бизнес-анализе. Самый яркий пример, который тут же приходит на ум, это когда у вас, например, есть предприятие и у вас есть некие внутренние данные деятельности этого предприятия, но вам, конечно же, важно знать и то, что происходит в мире. Открытые данные вам в этом должны помочь. То есть, вы анализируете данные из открытых источников и как эти данные соотносятся с внутренними данными предприятия. Вы не обязаны открывать данные своего предприятия (это коммерческие данные), но при этом у вас всегда есть возможность получить информацию извне и сравнить ее с внутренний информацией, и на основе этого сравнения принять важное решение по работе предприятия.

Следующий вариант использования открытых данных очень важный, и мы специально заострили на нем внимание в Школе открытых данных. Это журналистика данных. Кстати, следующее занятие в Школе как раз будет посвящено именно ей. Это отдельная междисциплинарная область знаний. В нее входят как поиск и предоставление цифровых историй, так и журналистское расследование на основе открытых данных, и другие не менее интересные вещи.

Еще одна область применения открытых данных - это научно-образовательная среда. Открытые данные предоставляют очень широкие возможности для научно-образовательных сообществ. Какие? Прежде всего, это предоставление открытых научных данных. Представьте, что у вас всегда под рукой находятся некие массивы научных открытых данных, с которыми вы можете проводить научные исследования прямо не выходя из дома. То есть для этого даже необязательно состоять в научных лабораториях. Ведь если у вас есть аналитический инструментарий по использованию открытых данных в научных исследованиях, вы можете уже на основе представленных источников вести научные исследования. Да, я понимаю, что это звучит, возможно, излишне оптимистично, но тем не менее я надеюсь, что к этому все придет, и это все будет. Тем более, что мировые тенденции в этой области существуют. И я думаю, что в других лекциях как раз уделю внимание именно открытым научным данным. Это глубокая и интересная тема, которая сейчас активно развивается на Западе, в том числе и в США.

Следующий момент — это взаимодействие граждан и органов исполнительной власти. Точнее, это то, как в идеале осуществляется это взаимодействие в гражданском обществе, когда власть прислушивается к гражданам, получает от них какой-то ответ и реагирует на него. На самом деле, это очень важный аспект взаимодействия, это часть той самой экосистемы открытых данных, одна из её важнейших составляющих - предоставление открытых  государственных данных.

**Перспективы**

Я подготовила вопросы для дальнейших размышлений <http://opendataschool.ru/2013/09/introduction-to-open-data/>, которые разместила в интернете и даже получила уже некоторый отклик. Я могу сразу сказать, что эти вопросы очень глобальные. Я сейчас зачитывать их не буду, потому что думаю, что вы их прочтете сами. Это, возможно,  несколько визионерские вопросы, но, тем не менее, я считаю, что открытые данные - это следующий шаг не только по работе в интернете, но это следующий шаг по работе в различных сферах, т. е. фактически многие профессии могут вполне поменять вектор своего развития. Яркий пример - это профессия библиотекаря. Казалось бы профессия, ранее не пользующаяся большой популярностью, но теперь, с появлением цифровых библиотек, цифровой библиотекарь становится очень важной и значимой профессией.

**Заключение**

В блоге “Школы открытых данных” я разместила <http://opendataschool.ru/2013/09/introduction-to-open-data/> полезные ссылки, касающиеся тем, затронутых в моей сегодняшней небольшой вступительной лекции. Там же выложены вопросы для дальнейших размышлений. Я бы с удовольствием вступила с вами в диалог по этим вопросам, потому что здесь важно понимать следующий аспект. Нельзя быть всезнайкой во всех областях знаний, тем более что перед нами открываются новые направления развития информационных технологий. Нужно постоянно учиться, нужно постоянно иметь в себе некий стимул не только познавать новое, но и никогда не бояться спрашивать, как именно оперировать данными, как их анализировать, какой использовать новый аналитический инструментарий. Это очень важно. Я могу сказать по своему опыту, что аналитического инструментария очень много. В основном, он представлен, конечно же, на английском языке и часто не переведен на русский язык, то есть придется так или иначе знакомиться с инструментами для работы с открытыми данными на английском языке. Это тоже надо иметь ввиду.

И основная моя мысль - в добрый путь! Я очень рада, что вы пришли, что у вас есть интерес к открытым данным, к работе с открытыми данными. Я думаю, что как раз на примере открытых данных, на примере новых областей знаний мы с вами совместно узнаем много нового. И я очень надеюсь на обратную связь от вас.

Сейчас я вам обрисовала некую направляющую стратегию освоения темы открытых данных. И в ней можно выделить 2 уровня. Один уровень, условно говоря, визионерский, это тема открытости и открытого доступа к данным, это отдельная тема для дискуссий, но тем не менее надо эту тему держать в голове; у нас есть открытая наука, которая тоже подразумевает обязательный открытый доступ к массивам научных данных; то есть то, что дает некий потенциал для проведения научных исследований вне лабораторий. А второй уровень - это тема взаимодействия с государством при помощи открытых государственных данных.

**Обсуждение**

**Вопрос из зала:** В чем смысл разграничения проприетарных и непроприетарных форматов данных? Почему в пятизвездочной модели они разведены на разные ступени?

**Ирина Радченко:** Существует такое понятие, как открытость — openness. Когда мы говорим про открытость, настоящую открытость, здесь надо понимать следующие вещи. Я приведу пример из журналистики данных, хотя то же касается и любой науки. Как многие из вас знают, для ученых в экспериментах всегда важна повторяемость, т. е. всегда должна быть возможность повторить эксперимент. Для того чтобы повторить эксперимент и получить достоверный результат, необходимо, чтобы у нас был доступ к исходным данным, т. е. в случае с журналистикой данных - к открытым данным. Чтобы доверять журналистской истории, нам нужно посмотреть, на чем основана была эта история, откуда был получен этот источник. Для этого необходимо, чтобы данные были выложены открыто, в формате открытых данных, чтобы каждый мог проверить и убедиться, что да, действительно здесь нету вранья, эти данные вполне достоверны и правдоподобные, им можно доверять, история изложена грамотно, и выводы сделаны верные.

**Комментарий из зала:** Если вопрос касался именно отличия CSV от Excel, то оно состоит в том, что в Excel может быть добавлено форматирование, т. е. это формат жесткий и из него нельзя выкинуть форматирование. Любой экселевский файл гарантированно дает возможность, например, шрифт изменить, еще что-то, - то есть вставить туда дополнительную информацию. А CSV, это просто текст через запятую, он всегда одинаково гарантированно читаемый и он не содержит ничего лишнего - шрифтов, колонок и т. д.

**Комментарий из зала:** Еще эта разграничение открытых и проприетарных форматов важно в применении к разработке программного обеспечения для работы с данными.

**Ирина Радченко:** Да, если мы говорим про открытость, у нас всегда должна быть предусмотрена возможность модифицировать программное обеспечение, и обязательно должна быть открытая лицензия, в том числе те лицензии по открытому программному коду, о которых я не упоминала в лекции, потому что я говорила именно о лицензиях, ориентированных прежде всего на базы данных. При этом обратите внимание, я даже не говорила про лицензию Creative Commons Zero - это лицензия, ориентированная именно на научные массивы данных (<http://pantonprinciples.org/>). В моей лекции были упомянуты лицензии, ориентированные именно на данные. Если мы говорим про лицензии, ориентированные на программное обеспечение, то это так называемые GNU-лицензии (<http://ru.wikipedia.org/wiki/GNU_General_Public_License>) и прочие. То есть для пользователя, который занимается той же журналистикой данных разница  между проприетарным форматом и открытым форматом данных, пожалуй, не столь существенна. Она скорее существенна с точки зрения разработки программного обеспечения.

Сейчас, кстати, еще идут споры в западных научных кругах, ориентированных на работу с данными (это Open Knowledge Foundation и прочие организации) о форматах, какой формат представления данных использовать: простой или более сложный. У любого формата есть свои плюсы и минусы. Более сложные форматы RDF и OWL, работающие уже с семантикой, подразумевают более высокий порог вхождения программистов и разработчиков, но при этом дают больше возможностей. Более простые форматы типа JSON или того же XML, с одной стороны, проще в использовании, но при этом они не предоставляют семантической компоненты, т. е. они выполняют другие задачи. Но с точки зрения доктрины открытости, конечно же, непроприетарный формат важен прежде всего потому, что у этих форматов есть больше возможностей и соответственно, если говорить глобально, более высокий потенциал достоверности. Всегда, когда есть возможность покопаться в исходном коде, есть возможность посмотреть, как работает система. Все программисты прекрасно знают, что всегда выше доверие к такому программному обеспечению. Но эта дискуссия уже на другую тему, потому что, как о проприетарного, так и о непроприетарного ПО есть тоже свои плюсы и минусы, с точки зрения формирования бизнес-моделей и т.д. Но это уже выходит за рамки нашей темы.

**Вопрос из зала:** А кто в России сейчас является драйверами открытых данных, т.е. кто двигает это направление?

**Ирина Радченко (указывая на Ивана Бегтина):** Ну вот вы сейчас видите драйвера. Еще есть, конечно же, Открытое Правительство.

**Иван Бегтин:** Давайте скажем честно, что все наши усилия они были бы бессмысленны, если правительство не пошло нам в какой-то момент навстречу. Как бы я ни ценил и ни любил себя самого, я скажу честно, что без Михаила Абызова (а открытые данные - это задача Открытого Правительства) у нас бы ничего не получилось и в том числе не появилось бы большого количества порталов, которые мы теперь можем использовать.

**Ирина Радченко:** Я думаю еще, что это все-таки комплексная вещь. Отчасти я надеюсь на такой подход “говори-говори, и когда-нибудь во рту станет слаще даже от слова халва”. Есть ведь множество замечательных инициатив, в том числе и у Минэкономразвития, которые дали толчок именно для того, чтобы министерства и различные российские органы государственного управления выкладывали наборы данных. Мы все видим, как это происходит, мы можем за этим наблюдать, и использовать эти данные. Мы даже проводили Экспедицию данных в рамках нашей импровизированной школы дата-журналистики, и мы убедились, что, действительно, по крайней мере некоторые наборы данных, выложенные Министерствами, вполне пригодны для обработки и на их основе можно делать визуализацию, и предоставлять различный материал для аналитиков. Это действительно так. И я предлагаю вам присоединиться к нашей компании, потому что у нас очень дружеская атмосфера, если какие-то вопросы есть, они разъясняются. И мы всегда рады всем участникам.

**Вопрос из зала:** Ирина, а как соотносятся открытые данные и авторские права?

**Ирина Радченко:** Это очень концептуальный и очень важный вопрос. Тут есть о чем подумать, и есть о чем подискутировать. Об авторском праве и четвертой части Гражданского Кодекса Российской Федерации. Это давнишняя и трепетная история. Вы понимаете прекрасно, что у нас нету готовых открытых лицензий в России. Это действительно важный момент, то есть организации могут выкладывать данные с открытыми лицензиями, и в основном они выкладывают данные с лицензиями Creative Commons.

Но еще раз: вопрос про авторское право - очень глубинный, важный и профессиональный. Он важен для всего этого движения открытости. Так что ответить на него сейчас в двух словах я не смогу.

**Вопрос из зала:** Вы сказали, что открытые данные в основном инициировались через Открытое правительство, которое было создано Медведевым, а если Медведева сменят, то что будет с открытыми данными?

**Ирина Радченко:** Можно я поступлю, как зануда? Смотрите, во-первых, это сказала не я. Во-вторых, я в своей лекции как раз и пыталась донести ту мысль, что это движение - это лавина, которую не остановить. Я уже даже не про Российскую Федерацию, я про именно перспективу и про работу с данными. Здесь основная мысль в следующем: если умеешь работать с данными — ты на коне. Если не умеешь работать с данными, то тут уже труднее. Умение работать с данными предоставляет большие и широкие возможности, и собственно те страны, которые будут до последнего сопротивляться этому движению, окажутся не у дел. Соответственно, с моей точки зрения, очевидно, что рано или поздно все равно все придет в одно русло, а русло в том, что работа с данными необходима как на уровне бизнеса, так и на уровне государственных органов. Здесь возможны варианты: либо жить в пещере с топором, либо все-таки сидеть за компьютером и уметь обрабатывать наборы данных.

**Иван Бегтин:** Я думал рассказать об этом в своей лекции, но раз возникает этот вопрос, давайте  все-таки отвечу с самого начала. Открытые данные появились до того, как заинтересовалось наше правительство, и во всех странах мира они появлялись до того, как ими интересовались местные правительства. В Великобритании, прежде чем появился местный портал, несколько лет активисты трансформировали госсайты в данные. Поэтому занимается этим государство или не занимается, мы этим будем заниматься в любом случае. Кстати, у нас на правительственном уровне все началось с того, что в 2012 году Владимир Путин подписал майские указы, в этом году Путин подписал собственно хартию по открытым данным. Я не думаю, что здесь есть хоть какие-то расхождения между какими-то ведомствами, которые сейчас собственно с данными работают. В этом смысле все действия находятся в общем ключе или в общем защитном ключе, т. е. какие-то там данные надо раскрывать, какие-то пока опасаются.

**Вопрос из зала:** Скажите, пожалуйста, насколько сейчас активны гражданские активисты и научные сообщества в плане предоставления данных?

**Ирина Радченко:** Зарубежные - очень активны. Постоянно проходят конференции, вебинары. Это очень ценная штука. Когда, даже не имея прямой возможности посетить какую-то конференцию, можно включить компьютер и поучаствовать во всем этом движении. Опять же, здесь важно понимать, что все конференции проходят на английском языке, везде важен именно английский язык, т. е. у тех людей, кто не знает английского, есть априори некий барьер. С этим я столкнулась, когда многие люди мне говорили: «Ирина, извини, раз оно на английском мы уже никак не можем поучаствовать». И в частности, это создает барьер между нашим научным сообществом и зарубежным. Хотя мы и стремимся к интеграции на уровне образовательных систем и на уровне научных систем, но пока, к сожалению, эта активность у нас в зачаточном состоянии. Я разговаривала с людьми из различных организаций, в том числе из Российской академии наук. Интерес большой есть, но пока нет четкого видения. И тут важно понимать, что все это очень ново (не только для России, но и для зарубежных коллег тоже это достаточно ново). Большинство основополагающих книг, в том числе те, которые я отразила в своем списке <http://opendataschool.ru/2013/09/introduction-to-open-data/>, вышли совсем недавно.

Например, замечательная книга, грозящая стать своего рода "технологической библией", - это книга Питера Субера “Open Access” <http://mitpress.mit.edu/books/open-access>. Она вышла совсем недавно в MIT-Press. Обязательно ее скачайте. Она выложена под свободной лицензией в открытом доступе. Автора зовут Питер Субер, с ним можно пообщаться, у него есть Twitter. Эта тема действительно весьма актуальна, в том числе и для Запада, но это все тема для отдельной дискуссии. В частности, по вопросу о публикации в открытом доступе научных работ, потому что это отдельная болезненная тема. Как мы все знаем, сейчас у нас для того чтобы получить доступ к научной статье нужно заплатить 30-60 долларов. Это очень много. Особенно учитывая постоянную потребность ученых в новейших научных статьях. Мало, кто может себе это позволить. И что важно, сами ученые при публикации из этих денег не получают фактически ничего. Эту тему сейчас уже потихоньку раскручивают, и есть уже варианты, как это все дело обойти.

**Вопрос из зала:** На каком этапе мы находимся, если говорить именно о связанных открытых данных?

**Ирина Радченко:** В мире действительно идет большое движение по связанным открытым данным, особенно в рамках образовательных центров и европейской программы рамочных проектов. Это все лежит в открытом доступе, можно посмотреть бюджеты, огромнейшие бюджеты, выделяющиеся на разработки по семантическим технологиям, по связанным открытым данным. У нас сейчас, насколько я знаю, присутствует интерес в различных организациях, у венчурных фондов.

**Вопрос из зала:** А насколько правительство заинтересовано именно в этом?

**Ирина Радченко:** Я думаю, что Правительство очень заинтересовано. На заседании Открытого Правительства в июле этого года были приглашены эксперты из сообществ, занимающихся именно семантическими технологиями. Это Зорен Ауэр из Лейпцигского университета и его коллеги из других стран, которые как раз занимаются проектами в области связанных открытых данных, в том числе, в рамках европейских программ с большим объемом финансирования.

**Вопрос из зала:** А Яндекс например?

**Ирина Радченко:** Я думаю, что у нас в Школе выступит специалист из Яндекса, так что, пожалуй, лучше узнать у первоисточника при возможности.

**Из зала:** Если первоисточник приходит сюда, значит и Яндексу это интересно. Можно последний вопрос, такой, резюмирующий, наверное. Я правильно понял, что в идеальном мире источниками открытых данных являются государственные структуры, общественные организации, научные организации, и все данные, которые порождаются в рамках этого процесса, должны быть общедоступными.  А как насчет коммерческих источников?

**Ирина Радченко:** Коммерческие организации тоже могут выступать источниками открытых данных, если они хотят заработать на этом деньги. Это тоже возможный вариант. Есть соответствующие бизнес-модели. Есть бизнес-модель предоставления данных в качестве сервиса (Data-as-a-Service). Есть различные бизнес-модели зарабатывания на данных, но это отдельная тема для дискуссии, и достаточно интересная. У меня есть ссылки на отчеты различных аналитических агентств (Deloitte, Gartner). Они на моем сайте выложены в разделе Data Science:  <http://iradche.ru/refs/data-science/>. Можно посмотреть аналитику именно по использованию данных в бизнесе и способах зарабатывания на данных.

...

А с предоставлением ведомственных данных все очень просто. Основной принцип такой: вы платите налоги, и вы имеете право знать, куда эти налоги идут. Вы имеете право знать, какие научные исследования вы финансируете. Собственно, с этой целью в этом году Обама и подписал меморандум об открытии государственных научных данных (<http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>). Я об этом в ознакомительной части не сказала, но в общем-то, правду не скроешь, и она всегда выходит наружу. И конечно же, этот меморандум Обамы был ориентирован, прежде всего, на предоставление научных данных американским гражданам, но как вы понимаете, интернет - явление международное, и соответственно доступ к этим данным может быть у любого человека, который умеет пользоваться интернетом. И основная мысль у меня следующая: Обама уже такой меморандум подписал, т.е. движение в этом направлении уже пошло.

Есть также замечательные примеры того, как в лабораториях и в домашних условиях использовались научные данные и получились отличнейшие исследования. Было несколько известных нашумевших историй, когда студенты сделали научные открытия и стали затем вестниками, сподвижниками идеи открытого доступа (Open Access): <http://novanews19.wordpress.com/2013/07/07/scientific-discoveries-open-access-and-jack-andraka-the-world-is-your-oyster-jack/>, <http://figshare.com/blog/Open%20Access%20Is%20Not%20Just%20For%20Scientists.%20It's%20For%20Everyone./72>.

Но не надо воспринимать это так, будто такие вещи происходят только благодаря решению какого-то политического лидера, вроде того же Обамы. Это комплексное решение. Обычно подобного уровня решения, особенно на государственном уровне, принимаются в комплексе и инициируются сразу с нескольких сторон.

И этим обусловлен международный государственный интерес к этой теме. Здесь ситуация такая: если мы не научимся этому, научатся зарубежные наши контрагенты, так называемый “потенциальный враг”.

**Вопрос из зала:**  А как у нас сейчас в России законодательство решает вопрос об открытых данных и персональных данных?

**Ирина Радченко:** Открытые данные должны быть представлены в деперсонализированном виде, обязательно, это отдельно прописано, да. Это требование, потому что все нормативно-правовые акты должны учитывать предыдущие.

...

Да, наверное, самое интересное заключается в диалоге. Кстати, это один из ярких моментов. Обратите внимание, что когда говоришь один, можно говорить о чем угодно, о своем, идти к своей цели, но когда находишься в диалоге, всегда выясняется, что, в общем-то, цель может быть и не та, которую ты ставил изначально. И раз цели у нас общие, то мы потихонечку их корректируем, и какие-то новые знания получаем совместно. Здесь важен именно позитивный аспект - стремление к получению новых знаний.

Это я все к тому, что сейчас пока эта доктрина (доктрина открытости знаний) еще не выкристаллизовалась, но она уже просматривается даже в названии фонда Open Knowledge Foundation («Открытые знания»). В идеале, это открытые знания и доступ людей в интернете именно к знаниям. Сюда же подходит и открытое обучение, в том числе и наша дата-экспедиция. Сейчас мы тоже занимаемся открытым обучением и проведением дата-экспедиций. Правда, у нас пока маленькая аудитория.

**Часть 2**

**Иван Бегтин**

**Власть и раскрытие данных**

У меня тоже сначала будет вводная лекция, я тоже буду говорить про открытые данные. Мы с Ириной будем пересекаться в некоторых темах. Спасибо Ирине за очень интересную вводную лекцию, я, как всегда, узнаю что-то новое для себя. Но я расскажу о некоторых нюансах. И потом уже перейдем к практической части. Прежде чем начать, я вам просто расскажу  про свой последний опыт, сразу живое, с полей. Я вернулся из Краснодара. Там у меня было такое очень интересное краснодарское регулярное мероприятие. Раз в месяц, в пятницу, они собирают экспертов (иногда, правда, пятница оказывается на среду, вот как это было в прошлый раз, но, тем не менее, - это пятница и по формату, и по настроению). И там собственно департамент IT города собирает кучу экспертов, которые говорят про умный город.

Я рассказывал про открытые данные. Вопрос, который мне задали, и который сразу запомнился. Вопрос-предложение: «Давайте раскрывать маршруты автобусов в реальном времени». Первая реакция была, что террористы узнают о том, как эти автобусы будут двигаться, и при случае могут взорвать. Да, это, конечно, была реакция официальных лиц, при этом, после чего мы стали с ними разговаривать, и он говорит: но слушайте, если ему подождать 5-10 минут, он все равно узнает, он просто расписание посмотрит и все равно узнает. Автобусы там очень хорошо работают, я сам несколько раз подходил к остановке, и они приходят прямо по расписанию.

Вот такие вот сразу возникают, то есть аргументы из серии не было бы хуже. Другие аргументы я слышал из серии, что вот бы нам сделать избирательные участки с детальностью до дома. Вот есть какой-нибудь избирательный участок, вот у человека будет свой адрес, то есть улица и дом, и чтобы список был депутатов на муниципальном уровне, на региональном уровне, на федеральном уровне, кто куда идет, кому писать кляузы, письма и так далее. И говорят, что надо публиковать эту информацию как открытые данные. Тут уже возникает другой вопрос: что вы сделали, для того чтобы найти эту информацию? — вот ничего. То есть есть закон восьмой ФЗ, по которому можно запрашивать информацию, есть активисты в МИДе, которые это делают, т. е. например, просто специально вручную по нарисованным кругам рассчитывают по улицам, как в Австралии делали. Ну и в России.

Вот Ирина мне подсказывает, что надо понимать, что мы не уникальны. Вот, как бы нам ни хотелось, чтобы у России был, там, третий-пятый-седьмой путь, неважно, - мы не уникальны. Все наши проблемы с бюрократией, с чиновниками, с корпорациями, со всеми проектами, которые не дают данные не являются уникальными. Просто везде скрывают все по-разному, у нас свои нюансы, у них свои. Убеждения и все эти отмазки, которые используют люди, которые должны передавать информацию, идентичные. Вот мне присылали список из пятидесяти отмазок, которые используют британские чиновники, чтобы не давать данные. У них перелом произошел несколько лет назад. Еще в 2005 году на каждом сайте висел копирайт “собственность короны”, сейчас его нету.

Поэтому я хочу сказать, что все вот эти изменения происходят у нас. Я когда разговаривал с одним человеком, он говорил: «Ну это же война с ними, это же надо требовать, это конфликт». То есть многие бояться идти даже не на конфликт, а на то, чтобы прилагать собственные усилия. На этой жизнерадостной ноте я приступлю к своей презентации, она будет отчасти похожа на презентацию Ирины.

**Открытость как идеология**

Давайте пойдем с самого начала. Вообще открытые данные существуют, действительно, не в вакууме, не в воздухе. Вот Ирина описала одну эволюцию открытых данных, я опишу эволюцию другую, движение за открытость и свободу. Почти все из вас наверняка слышали про открытый исходный код. Этому явлению на самом деле десятки лет. Оно происходило из научных учреждений, от различных анархистов, людей, которые говорили, что весь исходный код должен быть открытым, из разных других сообществ. Огромное количество групп, так называемых “евангелистов”, которые говорят, что открытый исходный код должен быть всегда. Они под это подводят научную основу, социальную основу, идеологическую основу, даже бизнес-основу, что открытый исходный код повышает качество продуктов, и в итоге многие компании могут экономить на том, что им нет необходимости продлевать лицензии каждый год, например. И сопровождение ПО гораздо более выгодное. Много споров, но это существует.

Кроме открытого исходного кода, существуют открытые стандарты. Это то, о чем мы говорили: проприетарные и непроприетарные форматы. Точно так же есть открытые стандарты, почти все протоколы интернета. В принципе интернет не мог бы существовать без открытых стандартов, без IETF, без W3C, и множества других координаций, которые они публикуют, и каждый может реализовать это. Это не патентованные форматы, их любой может использовать. Это открытая наука, то, о чем Ирина говорила, я не буду останавливаться. Открытые знания. Опять же Ирина затронула эту тему. Открытое аппаратное обеспечение. Это тема, которая пока еще мало заметна, хотя все больше. Это Arduino <http://arduino.cc/>, огромное количество других устройств. То есть, вам фактически не нужно покупать патентованное железо, закрытое оборудование. Если вы придумаете что-нибудь, будь то баллоны, дельтапланы, самолетики, машинки с камерами и т. д., которые выполняют еще какую-то функцию, вы можете это собрать из открытого оборудования, которое стоит буквально копейки. Вы не переплачиваете за устройство. Огромное количество движения существует вокруг этого. Есть люди, которые делают открытые тракторы, буквально открытую сельхозтехнику. Есть такая тусовка в Польше, они занимаются этим; выживальщики, т. е. люди говорят, вот мы напишем вам не просто, как выжить, а как вручную, а потом в кузне делать трактор, движок, очень много всего.

Иначе говоря, это все идеология. Идеология глобальной открытости, согласно которой благодаря этой открытости мы можем создавать нечто большее. Открытое государство - это на самом деле одно из следствий. Открытые данные - это то, что можно поместить по смыслу между любым из этих понятий, но суть в том, что доступность информации баз данных в машиночитаемом виде - это точно такая же идеология, как все остальное. То есть, когда мы говорим, что, если государство создает какой либо реестр, если коммерческая компания, осуществляя те функции, которые необходимо регулировать, например, контроль за ее экологической активностью или контроль тарифов естественной монополии, собирают данные, то чтобы они там не просто это публиковали, а чтобы это было открытыми данными, чтобы мы могли это использовать, проанализировать, создавать на этом коммерческие продукты или просто осуществлять контроль за этими институтами.

Поэтому первый участок определения открытых данных (хотя вы увидите много разных других) - это понимание открытых данных как движения, а движение - это люди, поэтому, в первую очередь, это идеология, и на этой идеологии собственно мы всё и делаем сейчас.

**От данных к знаниям**



Я очень люблю этот слайд, про него могу рассказывать очень долго. Обычно я сразу после него задаю вопрос. Поднимите, пожалуйста, руки, кто из вас умеет готовить. Ну, половина зала есть, прекрасно. Я думаю, что половина зала поймет мои а-ля аналогии. Представьте себе, что у вас есть стол, который усыпан овощами, фруктами, яйцами, какими-нибудь другими продуктами, на которые вы смотрите и думаете: “я могу из этого сделать и это, и это, а вот это я уже ел (или ела) вчера, а вот это я хочу съесть сейчас”. Вот я готовлю редко, но люблю в принципе. У меня товарищ профессиональный такой кулинар по рыбе, вот он готовит всегда. Он обычно раскладывает на столе кучу всего, включая специи, и думает: что же мне такого сегодня сделать? И у него шкафчик еще есть, и там всегда всего много. Это данные. Вот все, что я описывал, ингредиенты, - это данные. Данные - это не то, что вы можете съесть - вы же не будете есть соль горстями, вы же не будете есть специи или сырое мясо. Так и данные невозможно употребить напрямую. Это ингредиенты, то есть это то, из чего готовится.

Следующая стадия - это информация. Это обычно то, что дают нам госорганы, другие учреждения и так далее, то есть это информация, когда какие-то первичные данные переупакованы, но они не доделаны, то есть они представлены в том формате, в котором их можно съесть, но получить от этого какое-то удовольствие затруднительно. Презентация - это когда уже эта информация представлена в том виде, что нам это в общем-то интересно, не только сытно, но и красиво, то есть не только желудок рад, но и глаза. И наконец, знания - это когда все это употреблено. Так вот, собственно, это одна из главных трансформаций: когда мы говорим, что вот эти данные становятся открытыми, что эти ингредиенты становятся доступны всем. Эти ингредиенты создаются на общественные деньги, а все, что создается на общественные деньги, мы имеем право получать. Не только те люди, которые имеют к ним эксклюзивный доступ, то есть. куриная ферма, созданная государством, должна предоставлять бесплатные яйца всем. Но это не очень хорошее сравнение, потому что это физический объект.

Но открытые данные в каком-то смысле - это цифровой социализм. В отличие от тех же куриных яиц, мяса и т. д., использовать их можно неограниченно, их можно скопировать в любом количестве и использовать. Никто не имеет права нас в этом ограничить, и поэтому данные - это ингредиенты. Если кто-то вас будет спрашивать, какую они вам приносят пользу, подумайте о том, какую пользу приносят ингредиенты: из них можно что-то приготовить.

**Восемь принципов открытых данных**

Это собственно такое расширенное определение того, что такое открытые данные.

В первую очередь, открытые данные отличаются от всех остальных **полнотой**, то есть. то, что опубликовано, является целостным: это вся информация, которая только есть. Живой пример. Правительство Москвы совсем недавно опубликовало рейтинг школ, где указан номер в рейтинге и название школы. Являются ли эти данные полными и открытыми? Не являются. Потому что хотя они и говорят, что учитывалось то-то и то-то, но никаких первичных данных по школе нету, никаких полных данных нету, сколько реально баллов получила каждая школа, на основе чего, как и что, как данные собирались, - не предоставлено. Полных данных нету. Они из того, что сделали, дали очень коротенькую выжимку, и теперь осталось гадать, потому что некоторые критерии на самом деле спорные. Для кого-то важно наличие дошкольного учреждения в школе, а кому-то оно не важно, а оно влияет на общий рейтинг; а может мне это неважно, то есть не влияет на мой личный общий рейтинг.

Второй момент - **первичность**, то есть данные должны предоставляться ровно в той форме, как они собираются. Приведу пример. Допустим, есть какой-то государственный термометр, вот он стоит на каком-нибудь здании и транслирует непрерывно информацию. У него есть специальный протокол, он его публикует в интернете, он транслируется. А теперь представьте себе, что данные через этот термометр будут проходить так, чтобы они за день накопились, а потом приходит человек и говорит: так, 16 градусов это мало, надо написать 22, чтобы народ не бунтовал, а то будет плохо.

Так вот, это данные, которые не являются первичными - это данные, которые прошли обработку. Шутка шуткой, термометр, конечно, касается в меньшей степени государства, а в большей степени коммерческих компаний, которые транслируют какую-то информацию по экологии. Первичные данные - это когда вот ровно от того датчика и ровно до нас нет никакого промежуточного звена, и те данные, которые были собраны, должны быть предоставлены. А если между ними есть звено, то мы не знаем, являются ли данные по загрязненности настоящими или нас обманывают. Это первичность.

Теперь **своевременность**. Данные должны публиковаться сразу, как они появились, то есть если какая-то статистика появилась, вы должны ее получить не через полгода, а желательно уже в следующем месяце. Если данные были размещены в каком-нибудь государственном реестре или в частном реестре, то вы должны, опять же, получить это не через год и не через какое-то время, а сразу же, то есть как только они возникли.

**Доступность** - это обеспечение возможности, чтобы в принципе вы могли бы их скачать и получить свободным образом.

**Пригодность к машинной обработке** - это то, о чем упоминала Ирина. CSV, XML, XLS форматы, много всего.

**Отсутствие дискриминации по доступу** - это значит, что не только пользователи Google Chrome должны иметь возможность скачать эти файлы, но и даже те, кто пострадали от Internet Explorer’а. Та же возможность их скачать должна быть и у всех остальных людей, у которых есть те или иные технические и личные ограничения.

**Отсутствие пропритарных форматов**. Эта часть пересекается с доступностью. Но, т. е. фактически, несмотря на то, что большинство занимают позицию соглашательства и публикуют, например, в форматах Excel, вообще это не рекомендуется, и все стараются постепенно от этого уходить и этого избегать. Лицензионная чистота, опять же, на всех уровнях: у вас не должно быть ограничений на использование данных в коммерческих или некоммерческих целях. Для этого используют лицензию Creative Commons и во всяких экзотических случаях придумывают свои. Некоторые страны типа США и Великобритания несколько лет назад написали: Используйте это как хотите, только если вы не из Ливии или другой страны террориста. Как будто их это останавливало. Но вот есть такие пункты, так что, если вы увидите их в какой-нибудь государственной лицензии, то как раз по этой причине.

**Почему все делают ЭТО?**

Ну давайте начнем с того, а почему все делают это, почему люди вообще этим занимаются. Первое — **государственная политика открытых систем**. Это к вопросу о том, почему это происходит у нас. Государства в большинстве стран мира, мы не исключение, принимают стратегию, в которой постоянно говорят красивые слова: прозрачность, подотчетность, открытость. Иногда, конечно, мы относимся к этому с некоторым сомнением. Я, честно говоря, в шоке до сих пор от некоторых шагов, которые предпринимаются там. Российская открытость по госзакупкам, например, при всех ее проблемах, превосходит разумные пределы иногда, потому что раскрывается сильно больше, чем принято раскрывать в мире. Степень раскрытости информации по госучреждениям сейчас на новых порталах тоже огромная, в некоторых случаях информация крайне доступная, в некоторых случаях информации, наоборот, недостаточно, но государственная политика открытости - это не пустое слово. То есть реально очень многие данные раскрываются, либо они уже существуют в машиночитаемом виде, либо их получение пока затруднено. Проблем очень много, но де-факто это существует и де-факто это декларируется.

Далее, **государственная политика инноваций**. Инновации, опять же, - это то, на что можно ссылаться, говоря про открытые данные. Открытые данные — это экономика знаний. Компании, которые используют их - это все компании из экономики знаний. Яндекс бы не существовал, не будь ни открытых данных и свободы доступа к информации. Если бы каждый сайт был закрыт паролем или доступ к нему был бы только по подписи (пришел расписался и получил доступ в сайт), то тогда бы поисковые системы не существовали, и крупнейшие новостные каналы-агрегаторы бы не существовали. Огромное количество информационного бизнеса просто бы исчезло. И государство, когда декларирует свою политику открытых данных, оно декларирует и политику инновационную, и вы это еще не раз услышите, в этом есть доля разумности, хотя пока примеров не так много, тем не менее, это очень правильно.

**Международные обязательства**. Мы, другие страны - все подписываем всякие декларации, всякие международные петиции, международные законы и международные хартии. Совсем недавно, в июне месяце, Владимир Путин подписал Хартию об открытых данных G8. Это вот такой шаг довольно-таки исторический, то есть фактически Россия продекларировала свою готовность к тому, чтобы открытые данные были по умолчанию. Притом что, скажем честно, в большинстве стран мира далеко необязательно международные обязательства исполняются, не только в России, а в принципе. Но это некоторая претензия на то, что государство к этому готово. Как правило, после того как первые лица это подписывают, возникает куча активистов внутри страны, которые говорят: ну мы же подписали, давайте, давайте откройте наконец-то! Собственно, отчасти это делаем и мы. Если там написано, что криминальную статистику надо открывать, то начинается: “Ура, ура, наконец-то Путин это подписал, давайте откроем криминальную статистику”. То есть это всегда некий повод сказать: «Вот есть некие международные обязательства, давайте их выполнять».

**Экономия бюджетных средств**. Открываем данные и говорим: «Бизнес, мы даем тебе данные по транспорту, в реальном времени автобусы ходят по нашему любимому городу. Давайте вы сделаете кучу предложений. У нас нет денег, мы нищие, у нас начался кризис. Давайте мы сэкономим бюджетные средства и не будем делать предложения для транспорта, а вы сделаете это сами». Так открытым текстом говорят в городе Барселона. Они говорят: “ у нас мало денег, мы бедный город, в Европе кризис, поэтому мы сделали портал открытых данных и пусть бизнес теперь сам с этим мучается”. А если граждане к ним пристают, они говорят: «Ну значит востребованность не такая, и бизнес не хочет это делать, вот идите к коммерческим компаниям, мы им все отдаем, а взятки с нас гладки». В этом есть определенная логика и не очень плохая, я бы даже сказал, политика, когда есть достаточное количество таких компаний.

Наконец, **борьба с коррупцией**, то, о чем я говорил: госзакупки и все остальное, раскрытые декларации чиновников. Это важная тема.

Вот это краткий, неполный список причин, почему это есть. Помимо того, что это прилично, т.е. ты приходишь и скачиваешь какой-то реестр не в виде скана или не получаешь его в комнатке с пыльными полками рядом с туалетом в каком-нибудь госведомстве с двух до четырех часов, а ты имеешь право скачать и сразу проанализировать, т. е. есть очень много отличий. Открытые данные — это некий уровень культуры взаимодействия учреждений (не только государственных) с людьми.

**Государственная политика открытости**

Ну давайте я кратко пробегусь, я не буду здесь много останавливаться, но чтобы вы знали, что такое есть. 112-ФЗ — «Закон об открытых данных», реальное его название более сложное. Суть в том, что там дается определение открытых данных, это изменение в 8-ФЗ, который там упомянут ниже, «Закона о свободе доступа к информации». Ну, его полное название тоже более замороченное. Я не юрист, я не буду вас мучить юридическими терминами, но суть в следующем. В 8-ФЗ описывается право на доступ к информации и как это происходит, в 112 ФЗ описывается то, что данные должны быть в машиночитаемом виде. Это очень важный шаг, как и то, что это наконец пробилось. Я бы сказал так, чтобы ни происходило в нашем государстве, вот это будет точно заслуга нашего правительства и всех наших министерств, которые за это отвечают, в частности Минэкономразвития. Это очень важный шаг.

94-ФЗ «О госзакупках». Вы его все знаете, сейчас он 44-ФЗ. Это на самом деле был самый радикальный закон. Я до сих пор не знаю, что случилось с депутатами в 2005 году, но они его приняли, и Путин его подписал и после этого все проекты открытости по госзакупкам, которые только существуют благодаря одной единственной причине, что в законе было прописано жестко, что вся информация в публичном доступе бесплатна и все это стало доступно в публичном доступе. Все, что мы там находили: латиницу, коррупционные конкурсы и т. д. - возможно, благодаря только этому. Если мы считаем, что мы вот просто догнали всех остальных, это не так. В большинстве стран мира это не публично. В большинстве стран мира есть порталы для поставщиков, в которых надо зарегистрироваться, и только поставщики имеют доступ к этой информации, большинство граждан не имеет. И точно также не имеют доступа к информации больших контрактов.

У нас Минобороны по бюджету раскрывает данные медленно. В Соединенных Штатах задержка примерно год официально, потому что там есть оговорка на случай, “если мы начнем какие-то военные действия (т. е. нападем на какую-нибудь очередную страну) лучше, чтобы они бы не знали, как мы к этому готовимся”. Я, как человек, который анализировал данные, в том числе и американские госзакупки, я просто видел эти конкретные требования в их гайдлайнах.

**Вопрос из зала:** А может ли это измениться?

**Иван Бегтин:** Я думаю, что менее открыто уже просто быть не может. Нет, там к сожалению, уже не получится, потому что практически вся система государственной машины как раз построена на этой открытости, то есть там нужны гораздо более фундаментальные изменения и это просто невозможно уже. То есть можно открыть больше, но меньше открывать уже будет трудно. Ну, есть там всякие распоряжения ФСТ и ФАС России, потому что государство не обязано это раскрывать само. Например, ФАС требует раскрытия информации от РЖД. ФАС требует раскрытия информации энергетических компаний и энергетических монополий о тарифах, которые публикуются в разных форматах. Ну там у них информация Excel о тарифах. Можно взять тарифы у Мосэнерго и сделать мониторинг. Сделать интересное приложение “узнай свой тариф” - и оно тебе будет показывать на основе открытых данных, которые государство не публикует само, но принуждает к открытию, исходя из защиты прав потребителей. Это тоже на самом деле очень важная тема, которую я здесь пока не затрагиваю. Это большая тема, о которой я потом расскажу подробнее.

**Международные обязательства и отношения**. Мы подписали Хартию G8. Мы не входим в Open Government Partnership, это важно заметить. Там была история, когда мы хотели, но не вошли, но мы собираемся. Кроме нас, туда еще не входит много интересных стран. Германия, и Франция не собираются, а много стран типа африканских собираются. Это на самом деле ни на что не влияет, т. е. мы раньше гадали, влияет ли это или не влияет - как еще одно обязательство. Если бы нас взяли, было бы хорошо, не взяли — тоже не плохо, у нас есть Хартия G8, и мы претендуем на высокие позиции в рейтинге открытых систем.

Есть порядка пяти рейтингов, где есть открытые данные, это рейтинг открытого бюджета, рейтинг Веб-индекс и т. д., а чиновники они любят жить по рейтингам. Вот Минфин говорит: “Мы там в десятке, а мы хотим быть в пятерке открытости”. Им говорят: «Открытые данные». Вот если хотите быть в пятерке, нужно, чтобы как-то с ними заниматься. Есть какие-то другие ведомства и довольно много всего другого, поэтому благодаря тому, что все-таки есть определенные амбиции на вот эти высокие позиции, в том числе открываются данные внутри страны.

**Лицензии**

Про открытые лицензии немножко Ирина рассказала, я добавлю немножко с другой стороны. Большинство открытых лицензий создаются не государственным органом, а НKO. Они создаются Creative Commons, у них специальное там название типа фонда, Open Data Commons. Они соответственно немножко отличаются, где-то только данные, где-то это в принципе любые цифровые произведения. Они необходимы для стандартизации прав на использование данных, т. е. на самом деле это стандартизация, чтобы не надо думать о том, как написать свой текст. Ты просто повесил лейбл Creative Commons и живешь, причем многие вешают его партизанским образом. На сайте правительства висит например лейбл Creative Commons. Вот зайдите туда и посмотрите.

Открытые лицензии официально адаптируется правительствами. Вот то, что я говорил, что во многих странах мира есть практика, что правительства берут лицензию Creative Commons и каким-то образом искажают и пишут, что мы там ни за что не отвечаем, если вы используете эти данные. Так британцы делают на самом деле. Вот данные они опубликовали, но никакой ответственности за них нету. У них ответственность только за процедуру, то, что они должны раскрываться. А ответственности за то, что вы используете данные, нету никакой, и они очень жестко придерживаются этой позиции. Зато есть оговорки, что, дескать, используйте для чего угодно только не из страны-террориста. Если вы не хотите нам навредить и взорвать вокзалы, пароходы и так далее, то используйте как хотите. Это вот у американцев так тоже принято.

**Вопрос из зала:** А за достоверность несут ответственность в России?

**Иван Бегтин:** Ну, я думаю, что этот вопрос обсудим чуть позже, но вообще это сейчас предмет споров. Фактически нет, а де-факто, когда это вскрывается, то это очень быстро исправляется.

Открытые лицензии государства. Я тоже коротко подбегу. Три ключевые инициативы: AusGOAL, NzGOAL, UKGLF — это все международные инициативы Австралии, Новой Зеландии и Великобритании, где они говорят не просто про лицензии, а еще предлагают выбор лицензий в зависимости от типа того, что вы публикуете, как этим оперировать и вообще что это такое. На самом деле, это в первую очередь фреймворки, включающие обучающие материалы. Это очень важно в плане, когда вы разговариваете с людьми, почему нужно это и как это делать системно. Вот системно нужно делать примерно вот так: заходите туда и смотрите, как это делают в других странах.

**Мировые центры компетенции**

Теперь кратко пробежимся по тому, что происходит в мире, куда имеет смысл заглянуть и что посетить. Первое — Open Data Institute (ODI) <http://www.theodi.org/>. Это центр компетенции по открытым данным, созданный британским правительством возглавляемый Тимом Бернерсом-Ли. Они проводят обучение по открытым данным, они публикуют очень много материалов и если вы хотите понимать, что такое открытые данные, то начать можно отсюда. Это требует знания английского, но, безусловно, ознакомиться с этим ресурсом необходимо, если вам нужно правильное определение, что это такое, как это продвигается в мире и то, что касается ключевых трендов.

Далее, это Open Knowledge Foundation <http://okfn.org/>. Собственно, с позавчерашнего дня я являюсь послом  Open Knowledge Foundation в России. Я как Саманта Смит, посол доброй воли. Но суть в том, что это некоммерческая организация, моя любимая некоммерческая организация, честно говоря, из всех, кто этим занимается. Там такие отмороженные активисты по открытым данным, которые говорят мы “very pragmatic”, то есть они очень прагматичны. Они всегда говорят: “мы очень хотим, чтобы всегда была открытость, и мы всегда договариваемся со всеми, чтобы это было так”. Они запустили огромное количество проектов, они, собственно, разработали лицензию Open Data Commons, они очень активны и бодры. В материалах, которые мы потом опубликуем и разместим, будет информация о списке рассылки на русском языке, и в принципе я очень рекомендую посмотреть все их ресурсы и все, что они делают, потому что там очень много интересного, как об открытых данных, так и о многом другом. В большинстве случаев, опять же, на английском, но всячески рекомендую. Они, кстати, сделали Школу данных (School of Data) <http://schoolofdata.org/>. В отличие от нас, Open Data School, они не проводят оффлайновых лекций, у них обучение происходит только онлайн. Если там будут учебные мероприятия, очень советую поучаствовать, потому что вот у нас 2 человека от России поучаствовали - Ирина Радченко и Анна Сакоян - в учебном курсе School of Data, и, насколько я знаю, впечатление хорошее.

**Sunlight Labs** <http://sunlightlabs.com/> — это довольно экзотическая организация в нашем понимании. Это ребята, которые занимаются политическим мониторингом. Они не привязаны к какой-то конкретной политической партии, они завязаны только на технологии. То есть до того, как Конгресс США стал публиковать открытые данные, до того, как Белый Дом стал публиковать открытые данные,  до того, как Сенат стал публиковать открытые данные - они выдрали все из их сайтов, преобразовали в базы данных, опубликовали под открытыми лицензиями и сделали API, а потом с этими API еще и соревнования провели для разработчиков: «А давайте мы сделаем самое классное приложение по контролю Конгресса». И у них там куча приложений появилось на основе того, что они делают. Например, по измерению работы конгрессменов в количестве слов, которые он говорит за одну сессию. Ну, то есть – это, в общем-то, развлекательные вещи. Но они центр компетенции безусловно, они очень классные. Но надо понимать, что у них в большей степени гражданский контроль - это то, что мы бы назвали контролем над деятельностью конкретных политиков, а не государства в целом.

**Международные проекты**

Вот Ирина упоминала. ООН публикует открытые данные. Огромный портал, там можно посмотреть, что-то поискать, скачать. Он не самый совершенный. Международная бюрократия - это гораздо хуже, чем все национальные бюрократии. Что бы вы ни думали и что бы вы ни говорили, это обычно какой-то треш, ужас, но при этом даже они этим занимаются, это вообще удивительно. И даже следующая международная организация Всемирный банк. Как ни странно, Всемирный банк поддерживает один вот этот портал, при чем на самом деле у них 3 разных портала: по Африке, общий портал по странам, портал по различным вопросам и т. д. Если вы хотите сравнить, например, Россию с какими-то другими странами, вам сюда. Заходите туда, выкачиваете различные индикаторы и смотрите: по уровню образования, по уровню занятости женщин в правительственных структурах и много, много, много другого, в зависимости от того, что вам более интересно.

**Дополнительные источники информации**

Далее, ну, собственно, «Хочу знать больше». Вот если вы интересуетесь открытыми данными не только теоретически, но и практически, рекомендую обратиться у уже упоминавшимся ресурсам Open Knowledge Foundation, Open Data Institute, Всемирный банк. Еще - я не упоминал здесь ранее - есть Opengovdata.ru, это наш первый портал открытых данных, который мы сейчас преобразовали в новостной ресурс. Это непосредственно открытыеданные.большоеправительство.рф. Я бы сказал, если вы хотите получать официальную информацию, это первый источник. Там публикуется непосредственно то, что делают наши органы власти. Мы в Opengovdata.ru публикуем все, не только про государство, а вообще до каких данных наши лапы дотянутся, т. е. вот дотянемся до данных Мосводоканала - ну, негосударственные данные, но открытые. И, наконец, ресурс по журналистике данных Datadrivenjournalism.ru - это то, чем занимаются у нас несколько активисток, собственно присутствующие здесь Ирина Радченко и Анна Сакоян.

Это не все, т. е., несомненно, ресурсов гораздо больше, но пока я этим ограничусь и перейду к мастер-классу.

**Мастер-класс: демонстрация работы с Хабом открытых данных**

Заметки Ивана Бегтина по мастер-классу в блоге Школы: <http://opendataschool.ru/2013/09/ckan-1-mc/>

Я начну тогда издалека, но сразу к делу. У нас есть такой движок под названием CKAN, о котором упоминала Ирина. Это движок с открытым исходным кодом. В отличии от огромного количества проприетарных продуктов. Мы в других лекциях более подробно расскажем о том, как порталы открытых данных делают своими руками и быстро или как их сделать, если вы делаете что-то большое, но суть в том, что есть движок CKAN, который создала как раз Open Knowledge Foundation. Он открытый, его можно скачать, развернуть, поставить и непосредственно использовать.

Соответственно, Хаб открытых данных, который мы разворачиваем, как раз работает на движке CKAN. Почему мы используем его? Первое - это открытый исходный код. Второе — мы не платим ни за что, ни за какие лицензии, мы можем развернуть это сами, мы можем объяснить, как разворачивать это другим. Open Knowledge Foundation обеспечивает его поддержку, вы можете открыть Ckan.org , и вы увидите его.

Мы какое-то время назад перевели его на русский язык и создали возможность для работы с русским интерфейсом. И открыли наш портал под названием Habgovdata.ru. Вот он вначале существовал под названием Opengovdata.ru, а теперь он переехал на Habgovdata.ru.

Для чего это нужно? На этом Хабе мы сейчас практически ежедневно загружаем какие-то новые открытые данные. Открытые данные бывают двух типов. Открытые данные, которые официально публикуются какими-то госорганами и находятся в каких-то информационных системах; и открытые данные, которые мы каким-то образом выдираем из разных госсайтов - так называемый веб-скрэпинг. Все, что публикуется, публикуется двумя способами. Один способ - это когда мы всё делаем вручную, то есть мы заполняем карточку каждого массива данных. И второй способ - это когда в принципе с открытыми данными мы работаем через специальное API.

Итак, Хаб открытых данных. Здесь <http://opendataschool.ru/2013/09/ckan-1-mc/> подробно описана процедура регистрации, как можно авторизоваться на портале, как можно зарегистрироваться и авторизоваться. Это очень простая процедура. Точно также как вы это делаете с любой социальной сетью, вы регистрируетесь на нем, указываете имя пользователя, полное имя, логин и пароль или email и пароль, т. е. вы авторизуетесь. Я уже здесь зарегистрирован и не буду вам показывать эту процедуру, но суть в том, что вы регистрируетесь, вам приходит подтверждение, и вы в дальнейшем можете войти. Заданием на этом мастер-классе будет потом проделать это дома самостоятельно, т. е. вам нужно будет зарегистрироваться, авторизоваться, загрузить какой-то массив данных и попробовать поработать с API.

Давайте я пока параллельно покажу, как у нас непосредственно идет работа с массивами данных. Все данные, которые у нас собраны, можно посмотреть через поисковую систему сайта, она имеет в несколько уровней, то есть там есть обычно 3 вида просмотра — это группы, организации и пакеты данных.

Навигация во всех них довольно просто организованная, то есть в случае групп мы просматриваем все по специально подобранным группам. У нас здесь нет еще пока картинок, но есть данные ЦБ РФ, образование, пока еще не заполненное, данные ЕГЭ, то есть статистика тоже по образованию, данные системы ЕСИМО, это государственная система по морским ресурсам, ЕМИСС, государственные финансы, архивы официальных микроблогов, открытые данные города Москвы и много, многое другое. Это какое-то количество групп, которые можно посмотреть и которые можно добавить.

Следующий формат — это организация. Здесь у нас 4 организации, из них по двум внесены организации и внесены уже данные. Это Информационная культура и вот неизвестная, то есть мы пока еще не знаем какая это организация вносит, но вносим опять же мы. И, наконец, пакеты данных, это вот непосредственно формат поиска, формат навигации. Я не буду подробно останавливаться на том, как с этим работать с этим, так как на самом деле найти информацию не сложно. Вам достаточно ввести какие-то ключевые слова для поиска, и вы сможете найти и увидеть там данные. Вы набираете “данные Москвы” и получаете список данных, которые опубликованы по Москве. Наберете слово “статистика” и получите все статистические данные, плюс там есть какие-то теги, по которым можно осуществлять навигацию, плюс есть различные другие данные.

Я начну сразу с того, как с данными работать. Использование портала заключается в двух возможностях. Первая — использовать его как потребитель. Вы находите там то, что вам нужно, выгружаете данные с какими-либо целями: программирование или же это будет журналистика данных, аналитика и т. д. Вы находите данные, смотрите их описание, выкачиваете и работаете с ними уже у себя локально. Второй формат - это то, чем занимаемся мы в основном: загружаем новые данные. Если вы являетесь представителем органов власти или являетесь активистом, который хочет заниматься или в принципе вы хотели бы просто разобраться, как это работает, то прежде, чем скачивать, имеет смысл понять, как это происходит загрузка.

В качестве примера мы можем добавить какие-либо данные, которые у нас есть. Давайте я попробую продублировать те данные, которые мы ранее опубликовали, и провести публикацию от начала до конца. Так, у нас массив данных, ранее собранный нами с сайта МЧС. Это организация МЧС, в которой мы извлекли с их сайта противопожарной безопасности. Так мы и пишем: Организации структуры МЧС России. Портал автоматически будет подбирать название ссылки, но мы его заменим: mchs-orgs-new. Заполняем кратко описание, как правило: “Данные выкачаны 21 августа 2013 года с портала МЧС России и преобразованы в открытые данные. Вот примерно так. Далее мы заполняем теги. В данном случае у нас нету никаких данных привязанных у году. В случае со статистикой, например, если данные за 2011 год, мы указываем 2011 год. А здесь мы указываем такие, как МЧС России, организации, госструктура. Можно добавлять какие-то дополнительные требования, типа пожарные части.

Далее мы подбираем лицензию, и это одна из главных головных болей: как правильно подобрать лицензию. Несмотря на то, что в России официально лицензий нету, то мы все, что собираем сами, преобразовывая продукт, публикуем под Open Data Commons. Эта лицензия здесь присутствует. Мы специально пока не удаляли все остальные, хотя может быть потом мы поубираем ненужные.

Организация - Информационная культура, и мы говорим, что сделано это общественностью, то есть что мы сразу публикуем это для всех. Далее выбираем “добавить данные”.

Итак, далее идет процесс добавления ресурсов, то есть есть описание массива данных, это карточка. Карточка, содержащая определенные метаданные. Метаданных может быть достаточно много, и они могут быть довольно-таки разные, но при этом файлов, которые там могут быть приложены, может быть довольно много и они могут быть разными. Вот конкретно вот эти файлы нами уже подготовлены.

Некоторое время назад мы все эти данные выгрузили с сайта МЧС, мы их преобразовали в формат CSV. Соответственно эти данные уже являются автоматически машиночитаемые, они отчасти являются открытыми, но они не являются официальными и они не являются первичными. То есть на самом деле они не всем критериям соответствуют, но при этом мы их все равно относим к открытым. Вот у нас есть специальный branches, мы их выгрузили на специальный сайт на GitHub.com, вы можете его выложить куда угодно и в принципе их можно даже закачать непосредственно на сайт. В данном случае я подбираю ссылку. Вот моя ссылка сейчас откроется.

В общем, пока ссылка пытается открыться, мы возьмем первичный файл как он есть, скопируем эту ссылку и теперь добавим непосредственно в описание ресурса. Вот здесь мы указываем ссылку на файл, мы говорим, что это branches.csv и что это данные по структурам МЧС.

Я не буду в данном случае заполнять подробно все поля, обычно здесь заполняется еще структура полей данного файла, как он есть, чтобы могли повторно это использовать, то есть какое поле, что обозначает и указываем формат, в данном случае у нас формат CSV.

Можно сохранить и добавить другой файл, то есть мы сделаем “сохранить и добавить другой файл”, при этом, если мы хотим знать, откуда файл был взят, то мы делаем еще ссылку на непосредственный источник этой информации. Вот он у нас, это репозиторий на Github'е, и мы говорим, что вот так будет выглядеть ссылка, это репозиторий с исходным кодом, открытые данные МЧС. Мы не указываем здесь формат, потому что это не совсем открытые данные. Это поясняющие приложения, метаданные, описание, то, что обычно будет происходить и далее.

И уже нажимаем далее дополнительная информация, то есть заполняем карточку метаданных и вот здесь я заполняю: “Иван Бегтин”, указываю свой адрес в Информационной культуре. Опять “Иван Бегтин” и снова адрес свой в “Информационной культуре”. Здесь мы начинаем, когда идет выгруз в стандартное поле, это на самом деле подборка то, что мы сейчас настроили по умолчанию. CKAN расширяемая система. Можно писать расширения специально своих плагины, все конечно с открытым исходным кодом. Так в большинстве стран написано, она становится в базовое развертывание, оно позволяет достаточной гибкости. Мы здесь заполняем поле, например, орган власти и мы напишем МЧС России. В принципе мы можем сделать специальную карточку, которую можно заполнять. Вот сейчас мы все это сохраняем.

Итак, мы указали код МЧС России и непосредственно уже сохраняем. Конец.

Нами с нуля, фактически от начала и до конца, подготовлено описание массива данных вручную, то есть это то, что мы предварительно где-то собрали и это использовали, и массив теперь описан и подготовлен. На одном из следующих занятий я вам буду уже рассказывать о том, как это происходит с нашими органами власти, потому что это отличается. Сейчас мы делаем вводные вещи и мы публикуем данные так, как мы считаем правильным, а к органам власти другие требования. Они будут публиковать там с требованием контактного лица, емейла и так далее, потому что обязаны будут за весь этот документооборот от лица, ответственного там. У нас пока требований нету, это то, что мы делаем на общественной инициативе, но это очень важно знать, как это делают в госорганах не только на CKAN, но и в принципе где бы то ни было.

Что теперь это нам дает? Предположим, что я какой-то человек извне. Вот здесь есть 2 варианта: исследовать это или попробовать загрузить. То есть если я нажимаю сразу на “загрузить”, я открываю этот файл, я могу его непосредственно сразу использовать. Или я нажимаю “предпросмотр”.

Файлы можно точно также загружать непосредственно в систему, там есть возможность загрузить файл. Мы его закачиваем к себе и потом загружаем туда и еще там можно работать тысячью разных способов. Теперь непосредственно, что с этим файлом можно делать. Помимо того, что мы можем посмотреть по навигации, поискать там его, найти, и его может использовать кто бы то ни было - любой разработчик, журналист, аналитик, - в своих целях, мы можем оперировать всей этой информацией автоматически (см. раздел API в заметках к мастер-классу <http://opendataschool.ru/2013/09/ckan-1-mc/>). CKAN - штука полностью открытая. Вот есть CKAN, есть решение DKAN (это так называемый Drupal-based CKAN), и есть еще порядка трех платформ, которые сделали государство Индии, государство США и так далее. И куча проприетарных решений. Все открытые отличаются тем, что, во-первых, там открытый исходный код, а во-вторых, все, что там содержится, легко можно вытащить. Вот все, что там доступно, доступно для использования.

Так, я вам сейчас покажу маленький кусочек секретной информации, которую не принято показывать, но можно показать. Это вот информация, которую мы загружали. Собственно я тут под своим аккаунтом огромное количество данных туда загружал. Зарегистрировался как Иван Бегтин и с помощью специального API я загружал, там какие-то данные создавал, удалял, создавал по новой и для этого предоставляется API ключ. Вообще-то он должен быть секретным, но не страшно показать, то есть мы под ним загрузим что-то еще и потом его сменим.

API ключ вы получаете каждый раз, когда регистрируетесь на портале, он уникален, он позволяет вам все то, что я сейчас делал вручную, сделать автоматически с помощью специальных программ. Такие программы есть, они приведены в этом документе <http://opendataschool.ru/2013/09/ckan-1-mc/>. Это означает, что если вы либо умеете программировать сами, либо у вас есть разработчики, вы можете любые данные, которые у вас есть, опубликовать на портале. То есть заполняете карточку, отправляете и они будут опубликованы. И вместо вот этой всей процедуры руками. 99% файлов всех массивов данных, которые у нас опубликованы на портале, загружено автоматически. Мною, какими-то другими активистами, просто вот взяли всю систему и туда шарахнули три с половиной тысячи массивов данных. Взяли всю систему ЕМИСС и туда - шарах - 3500 массивов данных. Взяли систему ЕСИМО - и 300 массивов данных. Москва - 1700. выгружаем из различных ресурсов и автоматически загружаем к нам на портал. Это пример того, как это происходит.

Но при этом, несмотря на то, что мы можем авторизовывать данные здесь есть еще несколько особенностей, которые я собственно вам и покажу.

В документе, который я дал, есть так называемый публичное API, открытое API, для которого вам не нужно авторизовываться, оно доступно всегда. Предположим, за что-то вы меня невзлюбите и захотите сделать свой портал открытых данных. Вот там Иван Бегтин что-то нехорошее там делает, а мы сделаем свой в своем городе Вологде, в своей губернии и т. д., неважно. Вы хотите свое. Ну или вам не нравится интерфейс CKANа, и вы хотите сделать иначе. CKAN предоставляет открытое API.

Да, может быть, нам стоило сказать с самого начала, что открытые данные - это штука техническая. Я буду очень много вводить терминов, и если у вас будут вопросы, обязательно мне их задавайте. API - это технический интерфейс для программистов для работы с теми или иными информационными системами. Web API - это, соответственно, всякие веб-сервисы, через которые работают просто через HTTP-протокол - точно также, как Веб, только для компьютеров. У CKANа то, что вы сейчас видите, вот эти крокозябры, - это список всех пакетов, которые загружены сейчас в портал и которые можно выгрузить полностью. То есть все, что содержится у нас в базе данных, все можно выгрузить и все можно с клонировать; более того, у нас есть несколько порталов. У нас есть еще один портал, который мы делали по полиции, у нас был портал по Петербургу, который мы отдельно делали, и был портал, который мы делали по Москве. Суть в том, что из этих маленьких порталов, которые мы делали, мы все выкачали просто и засунули целиком в один, благодаря тому, что такая возможность есть. Вот то, что вы видите, это тот самый технический вывод всех пакетов, которые есть на сайте.

Другой список, это, например, список всех групп, которые есть на сайте. Вот все группы, которые есть. Можно все увидеть и использовать. Далее, все, что вы видите здесь, - это весь список тегов, которые мы используем. Вот все теги, которые есть, по которым можно искать. Это означает, что помимо того, что сами открытые данные присутствуют, открытые данные публикуются открытым образом и даже работать с ними можно открытым образом. Вы можете сделать свое приложение, свой сайт, свой ресурс, который скопирует, например, все эти данные, которые мы собрали по Москве или соберем по Санкт-Петербургу в единый Хаб, и публиковать у себя. Или вы можете сделать свой портал и закачивать еще к нам периодически, просто вот ссылка на него для того, чтобы привлекать внимание разработчиков, потому что мы хотим сделать такую централизованную историю, чтобы все могли это делать. И когда у нас появится Федеральный портал, точно также мы сделаем возможность, чтобы мы оттуда все это вытащили, и всем было удобно с этим работать и искать.

Поэтому вот это первая некоторая особенность. Кроме того, по каждому из объектов на самом деле мы можем получить более подробную информацию. Мы получаем информацию по тегу, и сейчас вы увидите этот список всех пакетов. Все данные, которые у нас опубликованы, машиночитаемы, то есть могут использоваться не людьми, а компьютерами (отчасти это касается упоминавшегося связанного веба). Вот мы вводим тег 2011 и получаем в машиночитаемом виде, так называемый формат JSON, все данные с сайта ЕГЭ. Мы их можем выкачать и работать только с ними. А дальше просто робот учится. Мы говорим роботу: вот возьми по такому-то тегу все данные, потом разбери, если там формат CSV, ты его разбери, преобразуй и можно сразу создать свою базу данных на основе опубликованной нами информации, то есть практически, это возможность открытым образом работать с ней.

Ну, и, наконец, еще небольшая техническая вводная, то, что упоминала Ирина. Это, например, Linked Data, связанные данные. Они точно так же в CKANe предусмотрены по умолчанию и все, что там опубликовано у нас, доступно как связанные данные. Вот данные mchs-orgs, которые мы опубликовали. Мы добавляем точка и пишем RDF (mchs-orgs-new.rdf) и выкачиваем как RDF файл. Сейчас мы попробуем его чем-нибудь открыть.

А, не RDF, RDF сейчас у меня не откроется в браузере, но суть в том, что файл, который мы сейчас закачали - он еще один вариант М3. И фактически все данные, которые есть, публикуются там. Ну, это еще одна накладка. У нас сейчас не выводится полностью информация, но суть в том, что данные, которые опубликованы, все массивы данных существуют и в RDF формате, и в М3 формате, и их можно непосредственно сразу использовать вот в тех самых инструментах, связанных с открытыми данными.

У меня, пожалуй, всё в плане мастер-класса. Кто здесь из присутствующих умеет программировать? Ну, примерно половина. Вот для всех, кто умеет программировать, это то, что на бумажке вы получили и то, что сейчас мы опубликуем у нас на сайте школы, как воспользоваться API CKANа. Сделайте что угодно, то есть любую вещь, которая вам понравится. Вы можете там сделать пакеты каким-то другим образом, сделать маленький скриптик, что-то загрузить с помощью этого API и получить ключ, но попробуйте это использовать. Те, кто не умеет программировать, попробуйте непосредственно использовать и опубликовать какой-нибудь массив данных, тест - потом его можно будет удалить, но попробуйте его сделать. То есть понять на своей шкуре собственно, что происходит с тем, кто публикует. Точно так же попробовать использовать, найти какие-нибудь данные по интересующим вас тематикам.

**Ирина Радченко:** Можете посмотреть в хабе данные, интересующие вас, и потом мы сможем их использовать в лекциях по дата-журналистике. Возможно, имеет смысл с этого ракурса подойти к использованию дата-хаба.

**Иван Бегтин:** После того как мы сделали вводную по нашему веб-хабу, притом что он, конечно, прекрасен, и мы себя любим и, конечно, хвалим и говорим “приходите к нам и используйте наши данные, которые мы там собрали”, я вам хотел бы сказать, что работая с хабом, который предоставляем мы, вы получаете возможность работать на самом деле с любым CKAN-развернутым ресурсом, который у вас есть. Таков и главный мировой хаб, который поддерживается как раз Open Knowledge Foundation - Datahub.io. Это огромный архив, просто огромное количество данных, и там все то, что я описал: все те же операции с API, с регистрацией, с доступом, с публикацией. И все это делается со следующими целями. Первое — привлекать внимание разработчиков, публикуя данные на одной центральной точке, чтобы мы могли это использовать, а второе, если вы являетесь разработчиком самостоятельно, то найти то, что вас интересует, и то, что вы можете использовать в практических целях.

Более сложные уровни работы с порталами включают в себя сразу загрузку данных структурированно, то есть вы получаете данные непосредственно из базы данных. Не просто файлами выкачивать, а уже получать данные оттуда, с геоданными и многим, многим другим. Мой сеанс агитации за CKAN закончился. Как говорится, мне не стыдно агитировать, потому что это открытый код. Вы можете забыть про наш дата-хаб, обидиться на меня, так сказать невзлюбить за что-то и взглянуть самостоятельно и сказать, что мы самая главная планета открытых данных. Я скажу: Ура, ура, я не один.

Теперь давайте приступим к вопросам и ответам, какое-то время еще есть у нас.

**Ирина Радченко:** У меня появилась идея. Вы могли бы зайти в блог «Школы открытых данных» и написать свои предложения по тем массивам данных, которые стоит использовать, в том числе, для визуализации в наших последующих лекциях. У нас есть готовые задания, но если вы предложите свой интересный вариант, я думаю, это будет интересно и разумно. В этом случае мы вместе будем конструировать работу с данными. Так что такой вариант тоже возможен. Я думаю, это самое интересное.

**Иван Бегтин:** Да, я думаю, то, что на следующих лекциях (они будут и теоретические, и практические) мы будем переходить к конкретным вопросам. Если у вас будут какие-то вопросы сейчас, то я отвечу сейчас. Если есть вопросы, на которые я не могу ответить сейчас, я думаю, что мы с вами договоримся, и я отвечу вам на следующей лекции.

**Ирина Радченко:** Или в интернете.

**Вопрос из зала:** Расскажите, пожалуйста, подробнее про связанные данные, на каком основании производится это связывание?

**Иван Бегтин:** Связанные данные — это сложная тема. Она в принципе сложная, она очень интересная. Я бы сказал так, связанные данные - это сейчас область скорее научного проектирования, и они происходят из научной среды. Вот Ирина, она как раз человек, который в большей степени занимался Linked Data, чем я, честно говоря. Я разрабатывал свою онтологию, но это скорее игры. Реально онтологию делают ученые, занимающиеся биологией, медициной, исследованием различных прав собственности даже и так далее, поэтому я не отвечу на этот вопрос быстро.

Об этом мы постараемся позвать Зорена Ауэра либо кого-то из его коллег, и они уже расскажут об этом подробнее. Я думаю, что на одной из следующих лекциях скорей Ирина посвятит этому чуть-чуть больше времени.

**Ирина Радченко:** В ходе нашей дата-экспедиции мы столкнулись с тем, что за рубежом есть много различных каталогов данных и поэтому, если будет возможность и желание организовать свой собственный российский портал открытых данных, это только приветствуется! Это будет замечательно.

**Вопрос из зала:** Возможно ли движком CKANа заменять базы данных, например при создании сайта?

**Иван Бегтин:** Внутри у CKANа, конечно, и SQL, и какие-то еще другие варианты типа SQLDB, которые просто неудобно использовать в промышленном развертывании. Я бы сказал так, у нас есть несколько вариантов развертывания CKANа, у нас был вариант развертывания CKANа с нуля, к  которому мы в итоге пришли, это проще, чем оперировать, и более сложный вариант, когда мы делали давно еще портальчик, в котором было сверху специальное приложение на Django, а внизу CKAN, и он работал через API. Такие варианты тоже есть. Ну, кому что нравится. Если вы делаете сайт государственный, скорей всего вам придется делать над CKANом надстройку, потому что CKAN грубоват, он не дает достаточной гибкости. Он неудобен в плане, чтобы его легко менять, поэтому обычно он поставляется разработчиками в компьютере, хотя по-разному. Вот правительство Южной Австралии просто развернуло CKAN и забыло про все, а правительство Канады сделало CKAN, а над ним повесило Drupal и сказало: вот мы будем обращаться к нему. То есть это вопрос технический, и мы можем рассказать об этом подробнее на одной из лекции. Еще вопросы?

**Вопрос из зала:** Я правильно полагаю, что в лекциях предполагается два потока - для программистов и для журналистов?

**Иван Бегтин:** Мы вначале думали о том, чтобы их разделить, но мы столкнулись с тем, что почти все записавшиеся к нам на курс сказали, что хотят и то, и то, поэтому я думаю, что мы будем здесь варьировать через один или каким-то другим образом. На самом деле то, что касается журналистов, касается и программистов, а то, что касается программистов, касается журналистов, потому что журналисты разговаривают с программистом понятным языком, а программистам надо уметь немножко визуализировать информацию и знать не только о том, что внутри, но и то, что должно быть на выходе.

**Вопрос из зала:** Я заранее поигрался с CKANом и, насколько я понял, CKAN не импортирует данные, загружаемые в него, в базу данных.

**Иван Бегтин:** Это не совсем так. Есть несколько способов. Во-первых, API CKANа позволяет загружать файлы. Если там версия последняя, оно не очень стабильно, но там есть разные варианты. Есть специальная Data Storage, это просто вообще-то в саму базу данных сохраняет, но там не всегда удобно, с большими данными тяжело. Есть возможность просто загрузить туда файлы, и он будет хранить в своем хранилище, но мы поступаем иначе, если честно. У нас просто есть аккаунт на облаке, и мы загружаем туда файлы. Один раз просто на облако, а потом там специальными скриптами, ну, там в облаке есть перманентные ссылки, и мы просто указываем в CKANе любые массивы - в Git, в Github. Если у нас гигабайт файлов, их неудобно грузить в Скан, ну просто ресурсоемко. Мы их выгружаем отдельно и подключаем отдельно.

**Вопрос из зала:** Правильно ли я понимаю, что API CKAN работает с файлом целиком, то есть нет возможности обратиться к строке?

**Иван Бегтин:** Есть, называется API datastorer. Вот API datastorer не очень пока стабильный и, в принципе, надо сказать, когда мы говорим про CKAN, что CKAN не самый совершенный по работе с открытыми данными. Он открытый, он самый динамично развивающийся и более простой в плане экономии бюджетных средств, своих денег, даже можно сказать бесплатный, то есть если администрация города Вологды захочет себе сделать портал открытых данных, вот так вот, развернули Скан и запустили там. Вопрос только его сертификации. Но есть более совершенные инструменты. Есть Socrata, есть FreeBase, они очень крутые, и мы потом отдельно расскажем о них. Это не совсем открытые данные, но это в принципе работа с данными. Но они все коммерческие, они все закрытые, они все находятся в американских облаках, их невозможно использовать в наших госорганах, ну, то есть там есть куча ограничений.

Да, домашнее задание мы пришлем, но на самом деле вам не нужно ждать его в письменном виде, оно всё равно то, которое я описываю. Домашнее задание очень простое: если вы не умеете программировать, найдите интересный массив данных, найдите массив данных и попробуйте его опубликовать на платформе CKAN. Я даже не переживаю, если вы сделаете не на портале нашего Хаба, вы можете зайти на Datahub.io, то есть глобально и опубликовать там, если хотите сделать на английском языке. Если на русском, то можно опубликовать его на Hubofdata, и мы все эти тестовые поубираем.

Если вы умеете программировать, то сделайте любой скрипт. Открыто, с авторизацией - просто проверьте, как это сделать. Но любая альтернативна, если вы все-таки это хотите сделать, попробуйте развернуть CKAN как раз таки самостоятельно. Так что у нас домашнее задание разделяется на два типа: для разработчиков, то есть тех, кто готов покопаться внутри, и для тех, кто хочет немножко поизучать, как это работает. Еще вопросы?

Тогда всем огромное спасибо, что все-таки добрались к нам в эту ужасную, холодную, дождливую погоду. Я очень рад был вас видеть и очень надеюсь увидеть всех вас на следующем занятии.