

ПИШЕМ О ДАННЫХ

Ищем героев и
делаем интервью
на основе данных

Я журналист.
Зачем мне цифры?

Кто дал
данные?

Пишем о данных

Пособие «Пишем о данных» научит вас создавать истории, основанные на данных. Здесь соединяются искусство сторителлинга и основы работы с данными. После прочтения вы сможете самостоятельно искать и находить нужные вам данные, переводить их в машиночитаемый формат, анализировать их и понимать, о чем можно рассказать с их помощью. Кто бы вы ни были – журналист, пиарщик, или студент, вы сможете применить эти навыки, чтобы находить истории в данных и рассказывать о них своим читателям.

финансируемый



Содержание

Почему дата-журналистика важна?.....	6
“Я журналист. Зачем мне цифры?” Журналистика данных и ее скрытые возможности.....	10
“Идея - путеводная нить по данным” Формулируем гипотезу, с которой будем обращаться к данным.....	14
“Кто дал данные?”	16
Об источниках данных и о том, насколько можно им доверять	18
Данные для компьютеров	19
Официальные источники данных.....	23
Другие источники информации.....	25
Расширенный поиск.....	29
Законы о данных.....	32
Извлечение данных (скрейпинг).....	35
Импорт CSV в Excel	40
Проверка данных.....	43
Организация датасета	46
Чистка данных.....	49
Накладываем фильтр.....	55
Базовый анализ данных, часть 1	58
Базовый анализ данных, часть 2.....	64
Этические аспекты в области журналистики данных.....	68

Почему дата-журналистика важна?

Почему мы нуждаемся в журналистике данных? Ответ совершенно прост: для безопасности журналистов и журналистики. На первый взгляд, это может показаться преувеличенным высказыванием. Но журналистика данных основана на фактах и цифрах, которые говорят все как есть (Журналисты "Говорите как есть!"). Об этом писалось на флажках редакции "Der SPIEGEL" во время его основания в 1950 году. И по сей день это остается актуальным для СМИ.

Мы менее уязвимы, если журналистские тексты исследуются исключительно на основе данных. Также никто не сможет с легкостью обвинить нас в предвзятости, дисбалансе и необъективности. С другой стороны, источники данных, большинство из которых опубликовано самими правительствами и администрациями, служат основой для журналистских расследований, а также дают возможность указывать на недостатки, не нападая на прямую, не подвергая себя опасности, без необходимости сгибаться и переписывать саму историю или оставлять ее незаписанной на столе по соображениям совести. Работа журналистов становятся неопровержимым, когда используются статистика и базы данных. Содержание остается прежним: забота о выявлении больших недостатков. Точнее: "Говорите как есть!"

Некоторые журналистские сюжеты могут быть поняты и объяснены только посредством анализа, а иногда и визуализации данных. Связи между влиятельными людьми или организациями могли быть не раскрыты, смерть от наркополитики, которая могла бы остаться скрытой и экологическая политика, которая наносит ущерб нашему ландшафту, могла бы оставаться неизменной. Однако каждый из вышеперечисленных пунктов нашли решение на основе данных, полученных, проанализированных и предоставленных в распоряжение читателей журналистами. Данные могут быть простыми, как простая расчетная таблица, или сложными, как данные демографических исследований, но все эти истории стоит рассказывать.

Как и любое хорошее журналистское расследование, журналистика данных может занять много времени, потому что она проводится с осторожностью. В свою очередь, быстро развивающийся медиа-бизнес практически не имеет капитала. Во всем мире появляется все больше спонсоров и сторонников, которые поддерживают проекты журналистики данных. Но им нужно объяснить, почему то, что они делают, важно и правильно. Как вы объясните тем, кто не является медиа-экспертом, что их поддержка для журналистики данных очень важна? Как вы убедите коллег переходить в сферу журналистики данных? Как вы объясните редакторам, что очень важно обучение коллег в

этой области, а также реализовывать и выпускать исследовательские проекты на основе данных?

Эта книга объясняет простым языком, насколько журналистика данных важна и как нам им пользоваться для улучшения жизни в обществе, используя всю мощь СМИ.

Журналистика данных это не просто точные исследования на основе баз данных и архивов. Она помогает нам фильтровать бесконечные потоки информации. Больше журналистам не нужно гоняться за информацией, они просто должны ее проверять, правильно классифицировать и развивать со смыслом и пониманием. Благодаря журналистике данных мы можем обеспечивать аудиторию наиболее важной и актуальной информацией. Важно отметить, что журналистика данных не скрывает свои методы и представляет результаты таким образом, что их можно так же, как и в науке проверить репродукцией.

Но помимо объединяющей цели анализа и предоставления информации, журналистика данных предоставляет новые подходы к рассказыванию историй. Благодаря постоянно растущему количеству инструментов, методов и подходов рассказывания историй, журналистика данных может обеспечить читателей, интересующих их информацией.

И в этом нет никаких сомнений: журналистика, основанная на фактах - это будущее. Журналисты должны иметь опыт в этой сфере. Раньше вы получали информацию, общаясь с людьми в поездах, а сегодня – через социальные сети. Для журналиста важно изучать данные и выявлять, что может быть интересно для аудитории, а также держать это в перспективе, помогая людям видеть, где все сходится и что происходит в стране.

Журналистика данных также может быть ответом на PR. Когда компании публикуют новые показатели, демонстрирующие их работу, а политики рассказывают о том, как снижаются показатели безработицы и растет ВВП, журналисты должны понимать предоставленные цифры. Особенно в области пропаганды, цифры часто используются, чтобы передать ауру серьезности - даже если эти факты придуманы.

В обществе, где производительность измеряется эффективностью, а прогресс поддается количественной оценке, необходимо наблюдать тенденции и выявлять возможности. Для этого требуются журналисты, которые знают, как обращаться с данными и имеют понятие о цифрах, полагаясь на официальные интерпретации. Только так можно обеспечить независимую интерпретацию официальной информации, а самое главное найти истину в растущем количестве цифрового контента. В

век информации журналисты нужны для курирования, проверки, анализа, обобщения и визуализации данных. В этом контексте журналистика данных имеет большое значение для общества.

Способность анализировать и интерпретировать данные является важной частью современного инструментария для журналистов, а не самостоятельной дисциплиной. Речь идет о том, как важно хорошо составлять отчеты и рассказывать истории подходящим образом. Журналистика данных - это еще один способ подвергнуть сомнению мир и привлечь к ответственности. По мере увеличения объема данных, очень важно, чтобы каждый журналист был осведомлен о методах журналистики данных.

Сила журналистики данных заключается в получении информации, которую иначе было бы очень трудно найти или доказать. В идеале журналисты используют данные для выявления областей интереса или других интересных вещей. В целом данные могут служить гидом или подсказкой. Истории исследуются на цифрах, которые объясняют их смысл. Журналистика данных - это способ видеть вещи, которые мы не могли бы увидеть иначе. Множество данных поможет нам рассказывать истории, отвечать на вопросы и обеспечивать понимание жизни.

Эта книга полезный ресурс для тех, кто заинтересован в журналистике данных и хочет попробовать свои силы в этой сфере. Она была написана в рамках проекта Немецкого института устойчивого развития с IDEM при поддержке Федерального министерства экономического сотрудничества и развития Германии (BMZ) и школы данных Кыргызстана (Data School Kyrgyzstan).

Авторы 1-16 Уроков - Алтынай Мамбетова и Анастасия Валеева (Data School Kyrgyzstan), Автор Урок 17: Этические аспекты в области журналистики данных - Пандели Пани.

Эта книга предназначена для журналистов и преподавателей университетов - и всех других специалистов, которые заинтересованы в журналистике данных. Мы надеемся, что книга станет богатым информативным источником журналистики данных, а также сможет рассказать о важности этой сферы и методах ее пользования. Книга даст читателям первое представление о журналистике данных, и покажет, на что именно надо обращать внимание.

Вы можете поделиться книгой людьми, которые могут быть заинтересованы в ее прочтении.

Др. Ульрика Фишер-Бутмалойю

Институт IDEM

Март 2018

Урок 1

“Я журналист. Зачем мне цифры?”

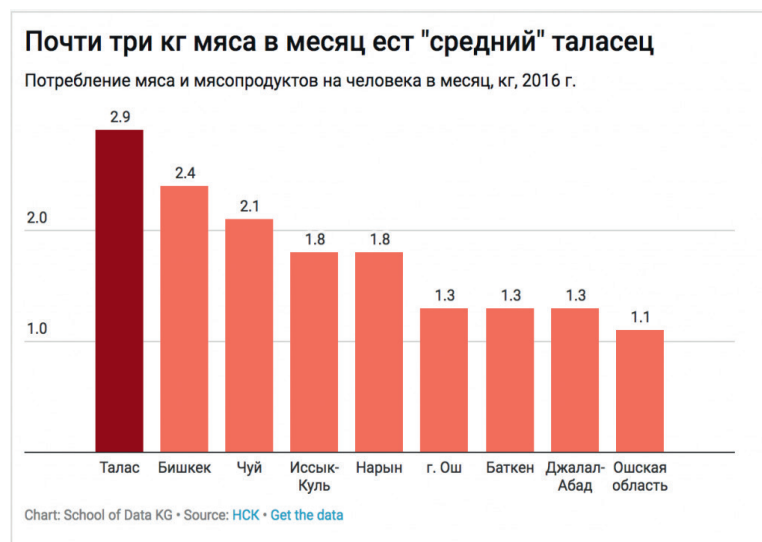
Журналистика данных и ее скрытые возможности

Короткое описание урока: Из этого урока вы узнаете о том, как и какие данные можно использовать в историях, в чем преимущество таких историй, и как их писать.

Привет, как дела?

Так начинается любой разговор. В ответ вы можете рассказать, как прошел день, перечислить, что у вас нового, поделиться наболевшим – словом, дать понять, что происходит в вашей жизни.

А как дела у нас в стране? Прямо сейчас, когда вы читаете эти строки. Начать этот разговор нам помогут данные. Если сейчас утро, страна просыпается, и в сотнях домов накрывают на стол, садятся завтракать. Какие здесь данные? Например, суточный объем калорий по областям и возрастам и количество потребляемых продуктов. Какие истории здесь прячутся? Например, в какой области люди недоедают, едим ли мы много мяса, и что вообще мы едим.



Эти и десятки других вопросов помогают понять, что происходит – не с отдельным человеком, а с группой людей, обществом, страной. Попытка это описать – это и есть сторителлинг на данных. С каких пор это важно?

- С тех пор, как человечество производит данные в большом объеме – со своих смартфонов, в интернете, да и просто в жизни, каждым своим действием генерируя социальную, гендерную, трудовую, статистику;
- С тех пор, как эта статистика по праву стала принадлежать не только государственным органам, но и гражданам в формате открытых данных;
- С тех пор, как нам есть, с чем эти данные сравнивать – другие страны и другие эпохи, стандарты и нормы, показатели Целей Устойчивого Развития ООН и другие.

С тех самых пор люди рассказывают не просто истории, а истории на данных. Например, такие истории, как эту:



Без вариантов. Почему школьники из бедных семей не могут сдать ОРТ и поступить в вуз

Работая над этой статьей, журналисты заметили некий тренд в данных: дети из бедных семей сдают ОРТ реже, чем дети из более обеспеченных семей. Журналисты задались вопросом: почему школьники из бедных семей не могут сдать ОРТ и поступить в ВУЗ? Попробуйте ответить на него сами. Вот несколько вариантов ответа, один из них - верный.

1. Они не сдают ОРТ, потому что собираются уехать на заработки за рубеж

2. При тестировании отбирают школьников из более богатых семей
3. В школе им приходится пропускать занятия из-за работы и у них не всегда есть учебники
4. По статистике, школьники из бедных семей не такие образованные

Прежде чем узнать правильный ответ, посмотрите на данные, которые использовали журналисты для подтверждения своей идеи.

- Результаты ОРТ по регионам
- Уровень бедности по регионам
- Количество школьников, не приступивших к занятиям, по регионам
- Денежные расходы на образование, по уровню дохода семьи

Эти данные они сочетали между собой. Например, совместив результаты ОРТ и уровень бедности по регионам, можно увидеть, что в наиболее «бедных» областях – это Ошская, Баткенская и Джалал-Абадская области – ОРТ не сдают чаще, чем в среднем по стране: почти каждый второй выпускник из бедных областей проваливает тест.



Как использованы в статье данные о количестве школьников, не приступивших к занятиям, и о денежных расходах на образование?

Таким образом, журналисты доказали, что школьникам из бедных семей приходится чаще пропускать занятия из-за работы и у них не всегда есть учебники. Поэтому и шансов сдать ОПТ на хороший балл у них меньше.

Наконец, когда журналисты доказали, что школьники из бедных семей имеют меньше шансов поступить в ВУЗ, можно предложить варианты, как изменить ситуацию.

Найдите варианты решения проблемы, которые предлагают журналисты.

Таким образом, журналисты рассказали историю о том, что сейчас происходит с образованием в стране. Получилась важная история, потому что они доказали, что у школьников из бедных семей меньше шансов изменить свою жизнь к лучшему, и образование не дает им дорогу в будущее. Данные были им нужны, чтобы: оставаться объективным, рассказать о целом явлении, а не об отдельно взятом случае и доказать систематическую несправедливость системы. А приемы сторителлинга – связный текст, наличие героя, визуализация – помогли донести основную мысль до читателя. Этому мы и будем учиться вместе с вами на протяжении курсов. Добро пожаловать в мир данных!

Урок 2

“Идея - путеводная нить по данным”

Формулируем гипотезу, с которой будем обращаться к данным

Короткое описание урока: Этот урок дает краткое введение в мир данных. Курс охватывает основные понятия и различные типы данных.

С помощью данных можно описать все вокруг нас. Но что это значит? Данные - это значения, присвоенные предмету. Возьмите, к примеру, банки меда на рисунке ниже.

Что мы можем о них сказать? Если это банки меда на продажу, то первое значение, которое приходит на ум - это цена. Но есть и еще.

Например, мед бывает разного сорта, это тоже даст нам много информации. Есть сезоны, есть место производства, вес, количество и качество. Все это - данные.

Из приведенного выше примера мы уже видим, что существуют разные типы данных. Двумя основными категориями являются **качественные** и **количественные** данные.

- Качественные данные - это все, что относится к качеству чего-либо: цвет, текстура и другие признаки, которые можно объединить в категории.
- Количественные данные - это те данные, которые выражены в числах. Например, вес меда, количество банок и т. д.

С данными удобно работать, когда они собраны и структурированы. Давайте сделаем это в таблице ниже.

Задание: определите, качественные или количественные данные в последней колонке.

Цвет	Белый	Качественные
Сорт	Горный	?
Вес в банке 0,5 литров	750 граммов	Количественные
Производство	Кыргызстан, Ат-Башы	?
Цена	200 сомов за банку	?

Но каждое отдельно взятое значение все еще не приносит нам пользы. Чтобы извлечь информацию из данных, их нужно интерпретировать.

Давайте возьмем вес: то, что пол-литровая банка весит 750 грамм, не говорит нам много. Но это имеет смысл, когда мы сравниваем вес и объем одной банки с другими. К примеру, можно найти информацию о том, что пол-литра меда весит от 700 грамм. Хорошо, теперь мы знаем, что вес меда в рамках нормы. Это уже информация. Но это еще не знание. Знание создается, когда информация изучается, анализируется и приводит к чему-то.

Что такое открытые данные?

“Открытые данные могут быть свободно использованы, модифицированы, и распространены кем угодно в любых целях”.

Это принятое в международном сообществе определение открытых данных максимально сжато формулирует суть открытых данных. Использованы, модифицированы и распространены – это три критерия, по которым данные считаются и могут быть признаны открытыми.

Данные могут быть **использованы**, если они доступны в интернете. **Модифицированы** они могут быть, если предоставлены в формате, который позволяет это сделать. И, наконец, **распространены** кем угодно в любых целях данные могут быть, если за ними не закреплены авторские права.

Таким образом, все три условия на самом деле технические – публикация данных, формат и лицензия определяют их открытость. Идеология прозрачности и движения за открытые данные началась именно с той позиции, что сбор этих данных уже был оплачен гражданами как налогоплательщиками, а, следовательно, доступ к ним - это такое же право каждого человека, как и другие фундаментальные права.

Больше информации об открытых данных [тут](#) и [тут](#).

Урок 3

“Кто дал данные?”

Короткое описание урока: В этом уроке мы будем рассуждать о персональных данных и о понятии data ownership, или владении данных.

Как правило, есть два исключения из открытых данных – это личные, персональные данные, которые охраняются законом о защите частной жизни, а также данные о национальной безопасности, раскрывать которые означало бы поставить под угрозу само государство.

Тем не менее, даже данные, которые изначально содержат персональную информацию, могут быть раскрыты, если информация по отдельным людям в них анонимизирована, то есть конкретные люди не могут быть идентифицированы. Другой вариант предоставления такой информации – агрегированные данные, то есть суммы значений, а не данные по каждому человеку.

Мои данные

Кто владеет информацией обо мне, кто контролирует ее, и кто имеет к ней доступ? Могу ли я получить копию данных о себе в таком формате, чтобы использовать или распространять эти данные, могу ли я извлечь из этих данных пользу для себя? Если я захочу открыто опубликовать некоторые данные о себе, могу ли я в принципе это сделать?



Пересечение открытых данных и моих данных

- Мои данные становятся открытыми данными (путем публикации): важные наборы данных, которые являются (или могли бы быть) открытыми, создаются из персональных данных посредством агрегирования, анонимизации и т.д. Значительная часть статистической информации — это опросы или статистика по отдельным людям, где конечные результаты агрегированы по группам (например, данные переписи или статистика рождаемости).
- Мои данные становятся открытыми данными (по моему собственному желанию): бывает, что люди хотят поделиться своими личными данными ради блага других. Пациент, страдающий онкологическим заболеванием, может поделиться историей болезни, если это поможет в поиске методов лечения ему или другим людям с этим же диагнозом.

Право выбора: если это мои данные, я должен иметь право на доступ, использование, распространение и раскрытие этой информации. Если открытые данные открыты для всех, то мои персональные данные должны быть открыты для меня.

Таким образом, персональные данные — важный источник информации, однако важно, чтобы активисты открытых данных осознавали риски, когда анонимизация и агрегирование данных не дает достаточной конфиденциальности. Еще более важно, чтобы правительство определило, как принимать решения о защите персональных данных и какие ресурсы понадобятся тем, кто будет публиковать эти данные. Важно четко понимать, где проходит черта между неприкосновенностью частной жизни и использованием открытых данных для общественного блага.

Урок 4

Об источниках данных и о том, насколько можно им доверять

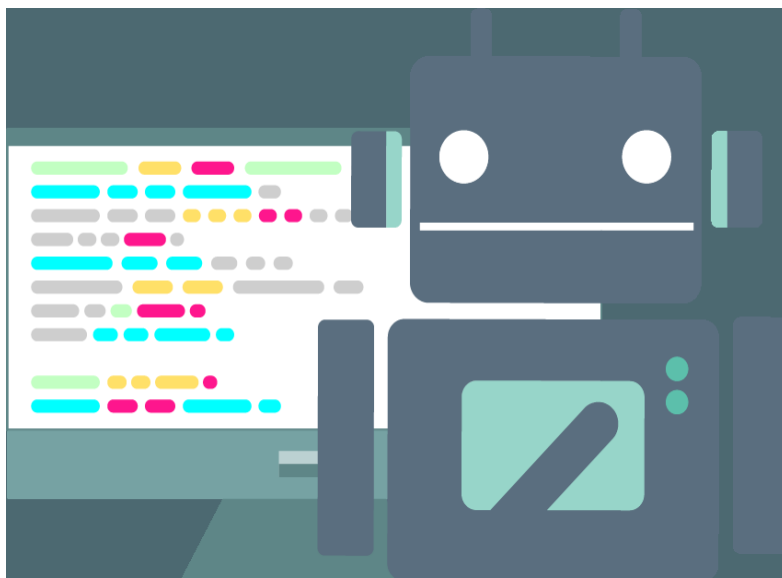
Короткое описание урока: На этом уроке вы узнаете о форматах данных для работы на компьютере.

Как появляются данные? Данные собираются в процессе **учета и регистрации** (например, данные о рождаемости), путем **анкетирования** (например, данные переписи населения), в процессе **голосования** (например, данные результатов выборов), в процессе **купли-продажи** (например, данные о продажах в интернет-магазине). Данные также образуются в результате работы мобильных устройств, датчиков, интернета, спутников (например, данные GPS) и многих других технологий.



Данные для людей

Смысл простого предложения «У нас есть 5 банок меда весом в 750 граммов по 200 сомов за каждую банку» легко понятен человеку, но не компьютеру. Предложение выше – это то, что мы называем неструктурированными (unstructured) данными. Такие данные не имеют фиксированной базовой структуры – предложение легко может быть изменено, и неясно, какое слово относится к какому значению. Аналогично, PDF-файлы и отсканированные изображения могут содержать информацию, которая понятна человеку, поскольку она хорошо написана, но не компьютеру, так как она не в машиночитаемом формате.



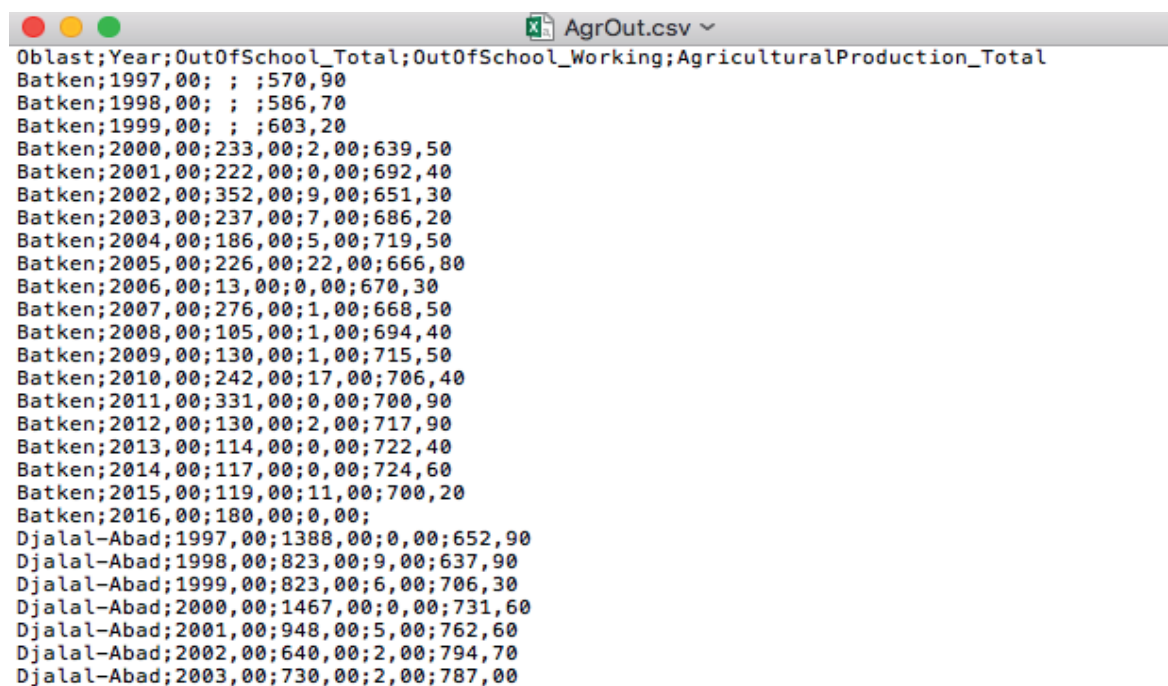
Данные для компьютеров

Компьютеры по своей сути отличаются от людей. Сложно заставить компьютеры извлекать информацию из определенных источников. Некоторые задачи, с которыми люди легко справляются, трудно автоматизировать с помощью компьютеров. Например, интерпретация текста, представленного как изображение, по-прежнему является проблемой для компьютера. Если вы хотите, чтобы ваш компьютер обрабатывал и анализировал ваши данные, он должен иметь возможность их читать и обрабатывать. Такая информация должна быть в структурированном и машиночитаемом формате.

Форматы данных: машиночитаемые, машинно-генерируемые, структурированные

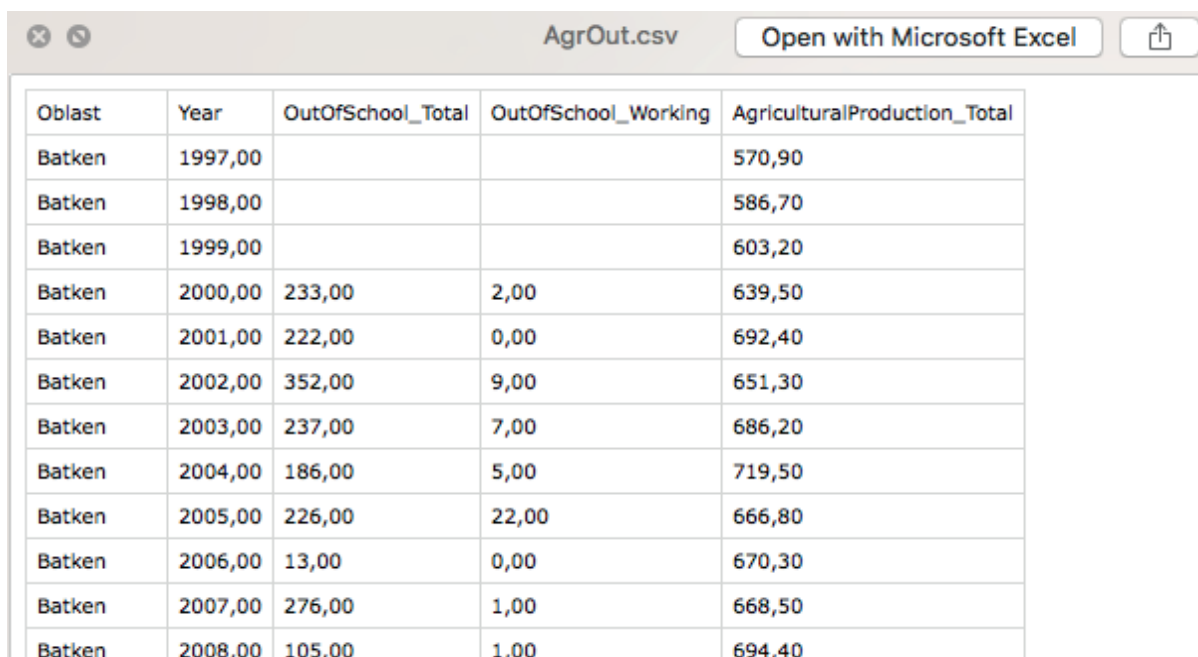
В этих форматах программное обеспечение может распознать структуру данных – чаще всего это таблица из столбцов и строк, которые организуют и описывают отдельные элементы данных. Самые популярные табличные форматы – Excel и CSV. CSV – это буквально «значения, разделенные запятыми» (comma-separated values).

Форматы CSV (значения, разделенные запятыми) и TSV (значения, разделенные знаком табуляции) – это форматы, которые используются для «кодировки» табличных данных. Проще говоря, файлы CSV и TSV – это текстовые файлы, в которых каждая линия – это строка данных, а колонки в ней разделены запятой (в CSV) или знаком табуляции (в TSV).



```
Oblast;Year;OutOfSchool_Total;OutOfSchool_Working;AgriculturalProduction_Total
Batken;1997,00; ; ;570,90
Batken;1998,00; ; ;586,70
Batken;1999,00; ; ;603,20
Batken;2000,00;233,00;2,00;639,50
Batken;2001,00;222,00;0,00;692,40
Batken;2002,00;352,00;9,00;651,30
Batken;2003,00;237,00;7,00;686,20
Batken;2004,00;186,00;5,00;719,50
Batken;2005,00;226,00;22,00;666,80
Batken;2006,00;13,00;0,00;670,30
Batken;2007,00;276,00;1,00;668,50
Batken;2008,00;105,00;1,00;694,40
Batken;2009,00;130,00;1,00;715,50
Batken;2010,00;242,00;17,00;706,40
Batken;2011,00;331,00;0,00;700,90
Batken;2012,00;130,00;2,00;717,90
Batken;2013,00;114,00;0,00;722,40
Batken;2014,00;117,00;0,00;724,60
Batken;2015,00;119,00;11,00;700,20
Batken;2016,00;180,00;0,00;
Djalal-Abad;1997,00;1388,00;0,00;652,90
Djalal-Abad;1998,00;823,00;9,00;637,90
Djalal-Abad;1999,00;823,00;6,00;706,30
Djalal-Abad;2000,00;1467,00;0,00;731,60
Djalal-Abad;2001,00;948,00;5,00;762,60
Djalal-Abad;2002,00;640,00;2,00;794,70
Djalal-Abad;2003,00;730,00;2,00;787,00
```

Так выглядят данные в формате **CSV**, открытые в текстовом редакторе



Oblast	Year	OutOfSchool_Total	OutOfSchool_Working	AgriculturalProduction_Total
Batken	1997,00			570,90
Batken	1998,00			586,70
Batken	1999,00			603,20
Batken	2000,00	233,00	2,00	639,50
Batken	2001,00	222,00	0,00	692,40
Batken	2002,00	352,00	9,00	651,30
Batken	2003,00	237,00	7,00	686,20
Batken	2004,00	186,00	5,00	719,50
Batken	2005,00	226,00	22,00	666,80
Batken	2006,00	13,00	0,00	670,30
Batken	2007,00	276,00	1,00	668,50
Batken	2008,00	105,00	1,00	694,40

Так выглядят те же данные, открытые в табличном редакторе

Стоит отметить, что есть еще много форматов, которые структурированы и машиночитаемы.

Подобные форматы, как правило, лучше всего подходят для анализа данных, и вы можете работать с ними в любой табличной программе, например, в Microsoft Office Excel. При поиске данных, если

вам удалось найти данные в формате Excel или CSV, это означает, что вам не придется тратить много времени на их форматирование.

Инструменты для работы с табличными данными: Libre Office, Google Таблицы или Microsoft Excel.

Данные в PDF

PDF файлы бывают нескольких различных видов.

Первый вопрос, который стоит задать при работе с PDF – созданы ли эти файлы на компьютере или нет? То есть, был ли файл сохранен в формате PDF, или его распечатали и отсканировали как изображение.

Второй вопрос – структурированы ли данные внутри файла? То есть, представлены ли они в виде таблицы со столбцами и строками.

Наконец, есть ли функция поиска по документу? Как правило, поиск возможен, если документ был создан на компьютере. Функция поиска позволит выделять текст внутри документа, и компьютер будет распознавать буквы и цифры.

Данные PDF в структурированном виде

Файлы PDF часто содержат структурированные таблицы, сгенерированные на компьютере, но формат PDF не подходит для работы с данными. Таблица должны быть преобразована в формат, который можно открыть с помощью табличных процессоров. То есть, требуется извлечь данные из PDF и перевести в подходящий формат с помощью специального программного обеспечения. Вы научитесь извлекать данные из PDF на практических занятиях курса.

Инструменты: Tabula, CometDoc, PDFtoExcel, Zamzar

Данные в отсканированных изображениях

Это, как правило, изображения, которые читаются компьютером как один гигантский блок информации, а не то, что можно разбить на отдельные части. Как правило, такие файлы изначально были сгенерированы на компьютере, а потом распечатаны и отсканированы. Таким образом, файл превратился в гигантское изображение. Для обработки таких файлов требуются программы оптического распознавания.

Примеры: Некоторые PDF и все растровые изображения (GIF, JPEG, PNG, BMP)

Инструменты: Google Docs OCR, Document Cloud

Данные в неструктурированных форматах

Некоторые данные были сгенерированы на компьютере, но не имеют структуры, распознаваемой компьютером. Это могут быть данные, записанные в виде текста, некоторые данные на веб-сайтах. В таком случае программист должен написать небольшой код, чтобы распознать структуру данных и извлечь их в необходимый формат.

Инструменты: языки программирования **Python** или **Ruby** для скрейпинга данных с помощью **morph.io**.

Менее распространенные форматы данных

Некоторые данные, особенно большие базы данных, сохраняются пакетами, которые можно отправить на веб-сайты или обработать с помощью статистического программного обеспечения, например, Stata или языка программирования для работы со статистикой R. Чтобы их можно было открыть в табличной программе, их нужно преобразовать в CSV или Excel формат.

Примеры: JSON (JavaScript Object Notation) или XML (расширяемый язык разметки) для программирования и .SAV или .R. Попробуйте использовать [Konklone](#), чтобы переконвертировать JSON файл в CSV.

Больше информации по форматам данных [тут](#).

Урок 5

Официальные источники данных

Короткое описание урока: Этот урок дает обзор официальных источников данных.

Основные способы получения данных из официальных источников:

1. **Поиск данных** - это поиск и обнаружение уже опубликованных данных.
2. **Скрейпинг данных** - иногда данные публикуются на веб-сайте, но без ссылки для загрузки. Эти данные можно получить с помощью "скрейпинга".
3. Получение данных по запросу (об этом в одном из следующих уроков).



Кто публикует данные для общественного использования?

5. **Правительство.** В последние годы все больше правительств выступают за открытие данных. Во многих странах создаются специальные правительственные порталы данных. Например, Нацстатком публикует много открытых данных. Подобные порталы данных существуют во многих других странах.
6. **Организации.** Другими источниками данных являются крупные организации. Например, Всемирный банк и Всемирная организация здравоохранения регулярно публикуют отчеты и наборы данных.

7. **Наука.** Многие научные проекты и учреждения генерируют свои собственные данные, некоторые из которых они публикуют для научного сообщества и широкой общественности.
8. **Это неполный список.** Данные могут быть опубликованы банками, корпорациями, мобильными операторами...

Чтобы помочь людям найти данные, были созданы списки баз открытых данных, такие как [Open Access Directory](#), [Open Data Inception](#) или [Datahub](#). Это библиотеки источников данных и отдельных наборов данных из разных источников.

Урок 6

Другие источники информации

Короткое описание урока: На этом уроке вы подробнее узнаете о других источниках данных, которые собираются с помощью НПО, краудсорсинга, журналистов и тд.

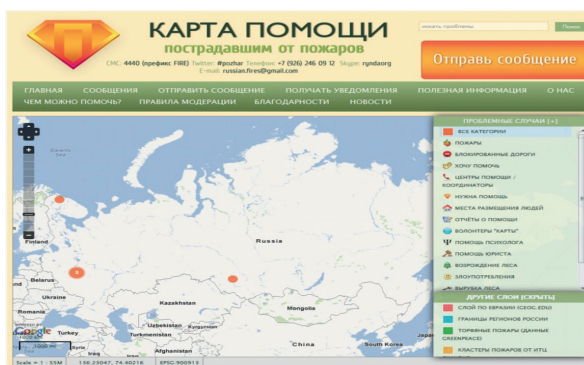
Часто, когда официальные источники данных недоступны, организации могут собирать данные методом краудсорсинга, с помощью граждан или группы подготовленных волонтеров. Другие альтернативные источники данных — базы данных неправительственных организаций (НПО), сенсоры, сообщения граждан, сообщения из средств массовой информации или утечки данных.

Данные, собранные методом краудсорсинга

Ushahidi — одна из наиболее известных платформ для сбора данных методом краудсорсинга. Ее используют для нанесения на карту подробной информации о природных катастрофах, политических кризисах и других событиях, где принципиально важен сбор и отображение данных в режиме реального времени. Система позволяет пользователям отправлять сообщения о происшествиях (таких как случаи насилия, фальсификации на выборах, преследовании со стороны полиции). Данные отправляются в централизованную систему для проверки, после чего база данных обновляется и сообщение привязывается к конкретному месту на карте.

Узнайте из этой презентации, как платформа Ушахиди была использована для создания проекта «Карта помощи»

Целью проекта был не мониторинг пожаров, а координация оказания помощи на основе базы данных о волонтерах и нуждающихся.



Данные гражданской журналистики

Во многих случаях средства массовой информации используют свидетельства очевидцев, а также призывают их присылать смс-сообщения, видео и фотографии — так называемая гражданская журналистика.

В этом примере радио Свобода пишет о сообщениях пользователей фейсбука: под тэгами #янебоюсьсказать и #янебоюсьсказать женщины и девушки делились рассказами о пережитом сексуальном насилии. Как правило, при публикации пользовательских сообщений СМИ должно их верифицировать, хотя никто не застрахован от публикации ложного контента.

Данные и дроны

СМИ все чаще используют беспилотные летательные аппараты для оценки размера протестов, объема свалок, уровня обезлесения, и для сбора другой полезной информации с помощью аэросъемки. Например, издание Wall Street Journal разместило видео, снятые дронами, чтобы продемонстрировать масштаб протестов в Гонконге.

<https://www.youtube.com/watch?v=qSes9FjdWbI>

Другой пример использования дронов – благодаря аэросъемке, активисты начали оценивать масштабы стихийных свалок в зеленых зонах Киева. Несмотря на доступность таких технологий, следует помнить о защите частной жизни и о том, что дроны могут вызвать у людей ощущение слежки или подозрение в том, что беспилотник запущен государственными органами.



Данные и сенсоры

Данные, собранные с помощью сенсоров, часто используются для репортажей об экологических проблемах. Например, СМИ может распространять небольшие, недорогие датчики среди волонтеров и обучить их ими пользоваться. С их помощью они будут собирать данные о качестве воздуха, температуре воды, или сейсмической активности.

Посмотрите на экспериментальный проект aba.kg Move Green по сбору информации о загрязненности воздуха в Бишкеке. Создатели проекта установили датчики в центре города, и любой горожанин в режиме реального времени может получать данные с трех точек в городе с оценкой загрязнения воздуха и рекомендациями по уменьшению последствий для своего здоровья.

Анализ данных СМИ

Когда официальная информация по теме практически отсутствует, сбор уже опубликованных сообщений СМИ может стать ценным источником данных. Такую стратегию применяли для сбора данных о насилии в отношении женщин, о погибших в перестрелках с полицией, и о помощи КНР странам Африки.

Например, проект «Файлы Мигрантов» собрал воедино все сообщения о мигрантах, погибших на пути в Европу. Данные были собраны из материалов СМИ и других доступных источников, например, баз данных негосударственных организаций. В совокупности, эти сообщения дали гораздо более полную картину гибели мигрантов.

Риски

Классическая проблема, связанная с использованием альтернативных источников данных — их достоверность и полнота. Например, если путем краудсорсинга собирать данные о неудовлетворительном доступе к государственным услугам, то вполне вероятно, многие из интересующих вас людей также ограничены в способах коммуникации, чтобы сообщить о проблеме. Вот еще проблемы данных, собранных методом краудсорсинга:

- Предвзятость выбора: информацию предоставляют только те, у кого есть время, ресурсы и мотивация.
- Процесс верификации: люди наводняют систему недостоверными данными, и проверить, какие факты соответствуют действительности, весьма сложно.
- Контекст: из таких данных мы узнаем только то, что сказали сами люди, то есть теряется контекст, который нужен, чтобы объяснить данные.
- Конфиденциальность: иногда личные данные участников проекта становятся доступны, таким образом, каждый из них может быть под угрозой.

Посмотрите на белорусский проект «Карта насилия», который поставил целью собрать данные о случаях домашнего насилия. Необходимо понимать, что поделиться такими данными могут далеко не все жертвы, соответственно, по собранным данным нельзя оценивать ситуацию.

Такие проблемы есть у каждого из методов сбора данных. Но это не означает, что с такими данными нельзя работать! Идеальных данных в любом случае не бывает. Важно оценивать риски и принимать во внимание ограничения данных при их анализе.

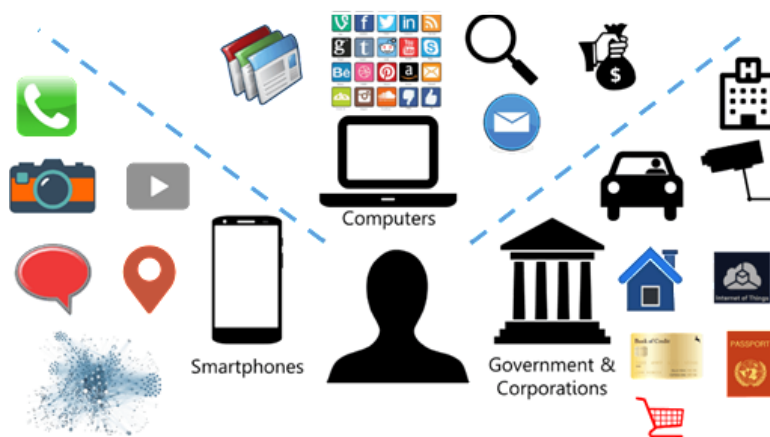
Урок 7

Расширенный поиск

Короткое описание урока: На этом уроке вы узнаете о том, как пользоваться расширенным поиском.

Сейчас, когда цифровые технологии развиваются очень быстро, нам доступно гораздо больше данных, чем когда-либо. Всё больше сайтов публикует данные и информацию, которая могла бы быть нам полезной, но найти её порой оказывается трудно.

В этом уроке мы рассмотрим способы поиска данных в интернете – как на специализированных сайтах, так и через обычный поиск в поисковой системе. Мы также обсудим, что делать, когда нужные нам данные недоступны, и как в этом случае их можно собирать при помощи краудсорсинга или сенсоров.



Расширенный поиск Google

В интернете находится множество источников данных. Удобный способ найти нужные данные онлайн — воспользоваться расширенным поиском Google.

1. Откройте расширенный поиск.

2. На экране появится окно с несколькими полями поиска.
3. Воспользуйтесь таблицей ниже: в ней описаны различные опции расширенного поиска Google. В левой колонке — описание различных полей в расширенном поиске. В правой колонке — эквивалент этих полей, которыми можно воспользоваться в обычной строке поиска.

Опции расширенного поиска в Google	Альтернативные варианты в обычном поиске
Со словами — работает как в обычном поиске Google	Введите все слова, которые вы хотите найти в обычной строке поиска
Со словосочетанием — помогает найти результаты, в которых слова находятся именно в том порядке, в котором вы указали	Используйте кавычки, например: "Министерство труда и развития человеческих ресурсов"
С любым из этих слов — помогает найти результаты, в которых встречается любое из введенных слов	Используйте OR между словами в поиске, например: сельское хозяйство OR фермерство OR урожай
Без слов — отфильтрует результаты поиска так, чтобы там не встречались указанные вами слова	Введите знак «минус» перед словом, которое вы хотите исключить из результатов поиска, например: Черногория -туризм
Искать на (язык) — показывает результаты на выбранном языке	—
Страна — ограничивает результаты до сайтов из определенной географической области	—
Дата обновления — ограничивает результаты недавним контентом	—
Сайт или домен: поиск на конкретном сайте	Используйте этот формат для поиска – site:url. Например: site:http://www.who.int/ или site:who.int
Формат файлов: Поиск файлов определенного формата (например: таблицы xls и csv, сканы pdf, документы doc)	Используйте этот формат для поиска – filetype:[расширение]. Например, так можно найти файлы с расширением XLS: БЕРНО filetype: xls НЕБЕРНО filetype:Excel

Задание: Попробуйте сами расширенный поиск Google.

- Используйте «с любым из слов», чтобы найти материалы о недоедании, голоде и истощении в вашей стране.
- Используйте «без слов», чтобы найти информацию о недоедании, но не о детях.
- Найдите материалы о гриппе только на украинском языке.
- Найдите материалы о гриппе только на молдавских сайтах
- Найдите материалы о гриппе за последнюю неделю.
- Найдите на сайте Министерства здравоохранения вашей страны файлы формата Excel
- Найдите PDF-документы о материнском здоровье в вашей стране.

Задание: Попробуйте найти файлы, используя строку поиска Google и ключевые слова.

- Отчет в формате PDF по образованию в вашей стране
- Отчет в формате PDF на веб-сайте ЮНИСЕФ по иммунизации в вашем регионе
- Годовой объем экспорта Вашей страны на www.tradingeconomics.com
- Файл Excel с данными о мигрантах из вашей страны
- Уровень инфляции в стране за последние 20 лет

Урок 8

Законы о данных

Короткое описание урока: Из этого урока вы узнаете о том, как правильно писать запросы в государственные органы для получения данных.

Если данных, которые вам нужны, нет в открытом доступе, их можно совершенно легально получить по закону о доступе к информации. Согласно этому закону, каждый имеет право доступа к информации государственных органов и органов местного самоуправления. Государство делает это для достижения «максимальной информационной открытости, гласности и прозрачности в деятельности государственных органов и органов местного самоуправления».

Получение данных по запросу об информации – это важный инструмент в получении данных! Во-первых, помните, на что у вас есть право. Во-вторых, пользуйтесь им. В-третьих, умейте аргументировать, если вам в нем отказывают.

Кому писать запрос

Запросы об информации могут быть направлены государственным органам и органам местного самоуправления в устной или письменной форме. Ответ на ваш запрос придет в той форме, в которой он был направлен: поэтому мы рекомендуем писать и отправлять запросы в виде документов. Тем не менее, никогда не будет лишним заранее позвонить в тот орган, куда вы пишете запрос, и уточнить определенные вопросы: собирают ли они данные по интересующему вас вопросу, по каким категориям и с какого года имеется информация, какова методология сбора данных и на чье имя отправлять запрос.

Как писать запрос

Давайте научимся правильно формулировать запрос. Допустим, вы хотите сделать материал о радиации.

Слабый запрос	Сильный запрос
Прошу предоставить информацию об уровне радиации в Мин-Куше	Прошу предоставить набор данных в формате Excel об уровне радиации в каждом из измеряемых объектов на территории Кыргызстана с 1990 по 2016 год, а также данные о допустимых нормах.

Задание: Вы делаете материал о неполных семьях. Напишите сильный запрос на данные.

Слабый запрос	Сильный запрос
Прошу предоставить данные о количестве матерей-одиночек в Баткене.	?

Помните, что вы не обязаны объяснять причину своего запроса – так гласит закон. Вы также не должны быть представителем СМИ или какой-либо организации, или объяснять, что вы будете делать с запрашиваемыми данными.

Запрос можно отправить и от своего имени, и от имени организации. Если вы журналист и готовите расследование, конечно, вы не обязаны в запросе на данные указывать, для какого материала нужны эти данные, это может оставаться внутренней информацией редакции.

Что делать после отправки запроса?

На запрос вам обязаны ответить в течение двух недель. Однако настоятельно рекомендуем сразу же по отправлении запроса связаться с отделом, который его получил, удостовериться, что запрос принят, и уточнить его входящий регистрационный номер.

По этому номеру вы сможете ссылаться на свой запрос в последующем общении. Кстати, если запрос перенаправляется в итоге в другой орган, то отсчет двух недель начинается заново. Поэтому

в ваших интересах во-первых, отправлять запросы с запасом времени, а во-вторых, заранее выяснить, какой именно государственный орган располагает этими данными. Также рекомендуем фиксировать всю переписку по вашему запросу.

Ответы на запросы граждан и организаций дают бесплатно, за исключением стоимости копировальных или почтовых услуг.

Что делать, если пришел отказ?

Вам не могут отказать просто так! Если вам прислали отказ в предоставлении информации, он должен содержать обоснование с указанием конкретных ссылок на нормы законодательства Кыргызской Республики, а также способы и порядок обжалования отказа в предоставлении информации.

Важно, что на некоторую государственную информацию установлено ограничение. Это информация, которая касается защиты национальной безопасности, охраны здоровья или нравственности населения, защиты прав и свобод других лиц. Как правило, это вся конфиденциальная информация, сведения о государственных секретах и охраняемых законом тайнах, оперативно-розыскной деятельности, а также информация персонального характера.

Однако знайте, что информацию персонального характера можно анонимизировать, чтобы сделать доступной. Именно по такому принципу публикуется статистика – без имен или ИНН, но со всей информацией, нужной для анализа: пол, возраст и регион, например. Если вы получили отказ на основании того, что это персональная информация, используйте этот аргумент, чтобы все-таки получить нужные вам данные. Мы проверяли – это работает.

Наконец, стоит помнить, что отказ в предоставлении информации или бездействие госорганов могут быть обжалованы по закону “Об основах административной деятельности и административных процедурах”. Иногда нам удавалось получить данные, напомнив об этом праве.

На заметку: если вы хотите знать все тонкости о вашем праве на информацию, ознакомьтесь с Законом Кыргызской Республики от 28 декабря 2006 года № 213 «О доступе к информации, находящейся в ведении государственных органов и органов местного самоуправления Кыргызской Республики».

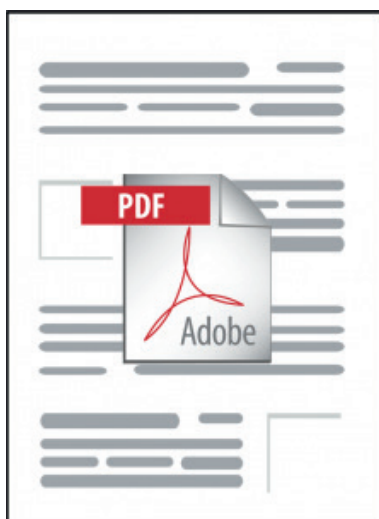
Урок 9

Извлечение данных (скрейпинг)

Короткое описание урока: этот урок научит вас конвертировать информацию в машиночитаемый формат.

Скрейпинг и техники извлечения данных позволяют создать таблицы из веб-страниц, PDF-документов или других немашиночитаемых форматов.

В этом уроке мы извлечем таблицу из PDF-документа.



Извлечение данных с помощью программы Табула

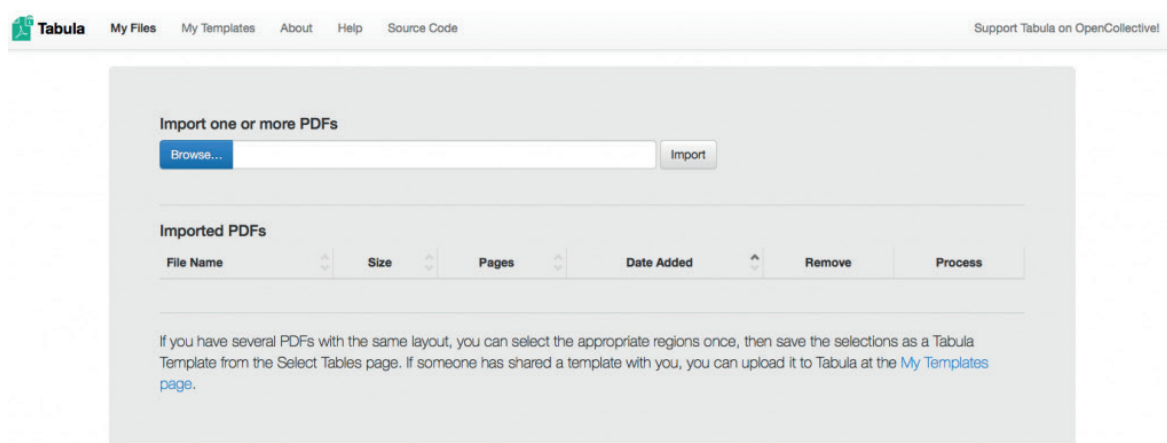
Табула – это программа для извлечения табличных данных из PDF файлов. Это очень удобный инструмент для большинства таких случаев. Программа открывается у вас в браузере, но использует локальное соединение – это значит, что при использовании Табулы интернет не нужен.

Установка Табулы

1. Убедитесь, что у вас на компьютере установлена Java, если у вас Windows. Ее можно загрузить по этой ссылке. Для MacOS этого не требуется.

2. Зайдите на официальный вебсайт Табулы.
3. Загрузите ту версию Табулы, которая подходит для вашей операционной системы.
4. На ваш компьютер начнет загружаться zip файл. Извлеките архив – вы получите папку «tabula».
5. Откройте эту папку и установите программу. Должно появиться всплывающее окно с кодом – это означает, что начался процесс установки.
6. После этого откроется окно веб-браузера. Это и есть Табула. Если браузер не открывается, наберите сами следующий адрес: `http://127.0.0.1:8080`

Так выглядит Табула в браузере:



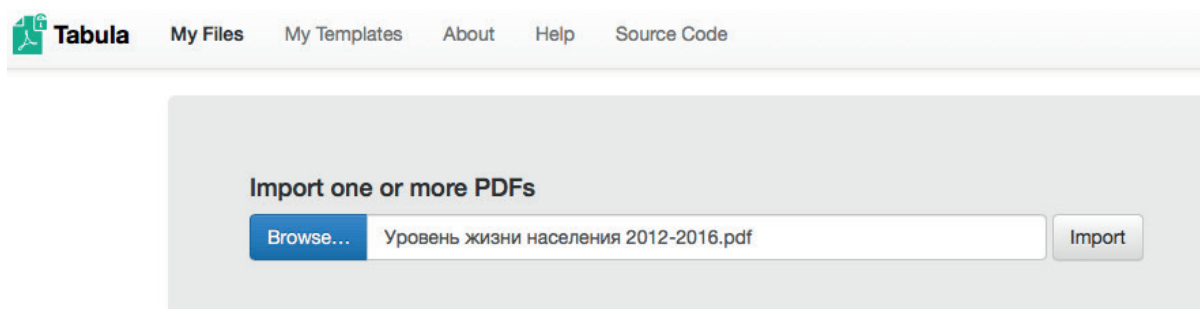
Пользуемся Табулой

Давайте воспользуемся Табулой, чтобы извлечь одну из таблиц из годовой публикации Нацстаткома [«Уровень жизни населения Кыргызской Республики 2012-2016»](#). Для начала найдите страницу 114 в публикации и таблицу “Таблица VI.н: Заболеваемость населения инфекционными и паразитарными болезнями”.

Давайте извлечем ее в табличный формат.

1. Когда Табула откроется у вас в браузере, нажмите кнопку Browse и найдите файл с публикацией.
2. Теперь нажмите **Import**. Это займет некоторое время, так как Табула будет подгружать весь

файл. Наконец, вам будет доступно превью файла.



3. Когда документ загрузится, прокрутите его до страницы 114, чтобы найти нужную вам таблицу.
4. Табула позволяет вам выбрать нужные таблицы внутри документа. Вы можете использовать функцию "Autoselect tables" для выделения всех таблиц, но при работе с большими документами лучше этого не делать. Вы всегда видите, что именно вы выбрали: эта область будет выделена красным. Выбор можно скорректировать, потянув за угол выбранной области. Функция 'Clear all selections' снимает все выделения.
5. Выберите таблицу "Таблица VI.н: Заболеваемость населения инфекционными и паразитарными болезнями".

Уровень жизни населения 2012-2016....

Templates Clear All Selections Autodetect Tables Preview & Export Extracted Data

Таблица V.1: Заболеваемость населения инфекционными и паразитарными болезнями (случаев)

	2012	2013	2014	2015	2016
Хронический тиф и паратиф	50	69	71	45	36
Корь	-	1	308	1783	3
Паротит эпидемический	377	690	892	569	328
Вторичный гепатит (исключая синдромный)	22226	20486	12024	11252	14615
Острые кишечные инфекции	21665	18466	18606	17894	18398
Астронтериты, колиты, вызванные					
установленными возбудителями	10444	9039	9819	9680	10777
Бактериальная дизентерия	1673	1540	1809	2131	1503
Тифы	2071	431	694	585	1713
Острые инфекции верхних дыхательных путей	191725	142012	159934	171640	261366
Сальмонеллезные инфекции	273	199	93	81	82
Дифтерия	1	-	-	-	1
Коклюш	63	94	113	280	1951
Мононуклеозная инфекция	8	12	28	90	16
Бруцеллез	2296	1364	1124	911	925
Бешенка	1310	1067	898	908	750
Детский	131	82	127	562	414
ВИЧ - инфекция ¹	704	480	612	605	714
Эпидемический	930	1049	1181	1134	963
Малярия	3	4	-	1	6
Сибирская язва	6	16	5	19	12
Бешенство	1	1	1	2	-

¹ Данные представлены по числу граждан Кыргызстана.

Repeat this Selection

- Важный момент: надо выделить всю таблицу так, чтобы выделенная область не включала в себя элементы, которые к таблице не относятся. Табула умеет распознавать табличные структуры в файлах pdf, и значит, выделять нужно только то, что внутри таблицы. В то же время важно не отрезать крайние ряды. Вы сможете сами это проверить через секунду.
- Нажмите **Preview & Export Extracted Data**.
- Появится следующее окно, в котором вы увидите превью извлеченных данных в структурированном, машиночитаемом виде. Проверьте, правильно ли выглядит эта таблица.

Tabula My Files My Templates About Help Source Code Support Tabula on OpenCollective!

Is the extracted data incorrect? You can revise your selected cells or try an alternate extraction method.

Revised Selected Cells Data has been extracted from the cells you selected in the previous step. You can revise your selection(s) to add or remove cells.

Choose Alternate Extraction Method The current preview uses the Stream extraction method. If the data is not mapped to the correct cells, try the Lattice method instead.

Stream Lattice

Stream looks for whitespace between columns, while Lattice looks for boundary lines between columns.

Still look wrong? Contact the developers and tell us what you tried to do that didn't work.

Уровень жизни населения 2012-2016.... Export Format: CSV Export Copy to Clipboard

Preview of Extracted Tabular Data

	2012	2013	2014	2015	2016
Брюшной тиф и паратифы	50	69	71	45	36
Корь	-	1	308	17783	-
Паротит эпидемический	377	690	892	569	328
Вирусный гепатит (включая сывороточный)	22226	20486	12024	11252	14615
Острые кишечные инфекции	21665	18466	18606	17894	18398
Гастроэнтериты, колиты, вызванные установленными возбудителями	10444	9039	9819	9680	10777
Бактериальная дизентерия	1673	1540	1809	2131	1503
Грипп	2071	431	694	585	1711
Острые инфекции верхних дыхательных путей	191725	142012	159934	171640	261366
Сальмонеллезные инфекции	273	199	93	81	82
Дифтерия	1	-	-	-	1
Коклюш	63	94	113	280	195
Менингококковая инфекция	8	12	28	90	16
Бруцеллез	2296	1364	1124	911	925
Чесотка	1310	1067	898	908	750

Если чего-то не хватает, вернитесь назад и измените выделение.

- Иногда заголовки отформатированы внутри таблицы в формате pdf таким образом, что когда мы их выделяем, Табула теряет структуру таблицы. В таких случаях, надо выделить таблицу без заголовков, а затем допечатать их вручную.
- Другой путь исправить вид извлеченных данных – поменять метод извлечения. Как вы видите из меню слева, есть два метода извлечения, **Stream** и **Lattice**. Stream ориентируется на пустые промежутки между данными, а Lattice – на границы ячеек. Попробуйте поменять метод извлечения, если вы не до конца довольны результатом.
- Наконец, когда таблица извлечена правильно, экспортируйте файл в формате CSV с помощью функции Export.
- Теперь с этими данными можно работать в табличной программе вроде Excel. Но сперва нам нужно импортировать CSV: об этом в следующем уроке.

Урок 10

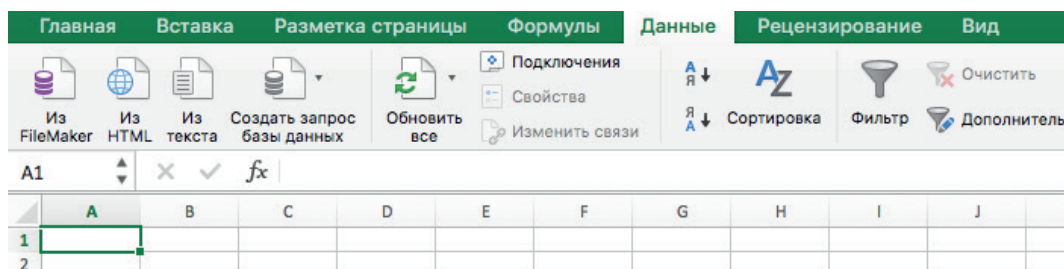
Импорт CSV в Excel

Короткое описание урока: в этом уроке вы научитесь импортировать файлы CSV в Эксель.

CSV нельзя открывать двойным щелчком. Проблема в том, что при таком открытии ваш файл не становится таблицей. Вы увидите текстовый файл с множеством строк и запятых. На самом деле, все эти значения – это и есть ваша таблица. Но чтобы её открыть, нужно импортировать данные из текстового CSV файла в Excel.

Давайте это и сделаем.

1. Создайте новый документ Excel.
2. Во вкладке Данные выберите «Из текста».



3. Найдите файл, который вы только что извлекли с помощью Табулы, и нажмите «получить данные».
4. В новом окне откроется мастер импорта текстов.
5. Выберите список значений «С разделителями». Таким образом мы даем Excel понять, что значения в нашей таблице разделены запятыми. Нажмите «Далее».

Мастер текстов (импорт) — шаг 1 из 3

Данные восприняты как список значений с разделителями.

Если это верно, нажмите кнопку "Далее", в противном случае укажите тип данных.

☒ С разделителями — значения полей отделяются знаками-разделителями
☐ Фиксированная ширина — поля имеют заданную ширину

Начать импорт со строки: Формат файла:

Предварительный просмотр выбранных данных:

Предварительный просмотр файла /Users/altynai/Downloads/tabula-2012-2016.csv.					
1	"	2012,2013,2014,2015,2016			
2	"	Брашной тиф и паратифы",50,69,71,45,36			
3	"	Корь",-1,308,17783,-			
4	"	Паротит эпидемический",377,690,892,569,328			
5	"	Вирусный гепатит (включая сывороточный)",22226,20486,12024,11252,14615			
6	"	Острые кишечные инфекции",21665,18466,18606,17894,18398			
7	"	Гастроэнтериты, колиты, вызванные",,,,,,			

Отмена < Назад Далее > Готово

6. В следующем окне, вы можете выбрать разделители. В нашей таблице – Знак табуляции и Запятая. Предварительный просмотр позволяет оценить, правилен ли ваш выбор. Нажмите «Далее».

Мастер текстов (импорт) — шаг 2 из 3

На этом экране можно указать разделители, содержащиеся в данных.

Разделители

☒ Знак табуляции
☐ Точка с запятой
☒ Запятая
☐ Пробел
☐ Другой:

☐ Считать последовательные разделители одним
 Ограничитель строк:

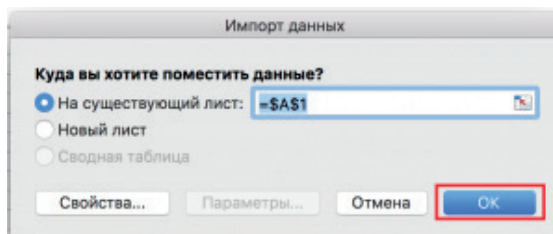
Предварительный просмотр выбранных данных:

	2012	2013	2014	2015	2016
Брашной тиф и паратифы	50	69	71	45	36
Корь	-	1	308	17783	-
Паротит эпидемический	377	690	892	569	328
Вирусный гепатит (включая сывороточный)	22226	20486	12024	11252	14615
Острые кишечные инфекции	21665	18466	18606	17894	18398
Гастроэнтериты, колиты, вызванные					

Отмена < Назад Далее > Готово

7. В третьем окне вы можете установить формат данных в каждой колонке. В данном случае, оставьте его как Общий, и нажмите «Готово».
8. В следующем окне выберите «На существующий лист», и проверьте, что первая ячейка это

A1 (если только вы не хотите поместить таблицу на новый лист в уже существующей таблице).



9. В результате ваши данные должны выглядеть вот так:

Главная Вставка Разметка страницы Формулы Данные					
Из FileMaker Из HTML Из текста Создать запрос базы данных Обновить все Подключения Свойства Изменить связи					
A1					
	A	B	C	D	E
1		2012	2013	2014	2015
2	Брюшной тиф и паратифы	50	69	71	45
3	Корь	-	1	308	17783
4	Паротит эпидемический	377	690	892	569
5	Вирусный гепатит (включая сывороточный)	22226	20486	12024	11252
6	Острые кишечные инфекции	21665	18466	18606	17894
7	Гастроэнтериты, колиты, вызванные				
8	установленными возбудителями	10444	9039	9819	9680
9	Бактериальная дизентерия	1673	1540	1809	2131
10	Грипп	2071	431	694	585
11	Острые инфекции верхних дыхательных путей	191725	142012	159934	171640
12	Сальмонеллезные инфекции	273	199	93	81
13	Дифтерия	1	-	-	-
14	Коклюш	63	94	113	280
15	Менингококковая инфекция	8	12	28	90
16	Бруцеллез	2296	1364	1124	911
17	Чесотка	1310	1067	898	908
18	Педикулез	131	82	127	562
19	ВИЧ - инфекция 1	704	480	612	605
20	Эхинококкоз	930	1049	1181	1134
21	Малярия	3	4	-	1

10. Сохраните файл с названием, по которому его можно будет легко найти в будущем.

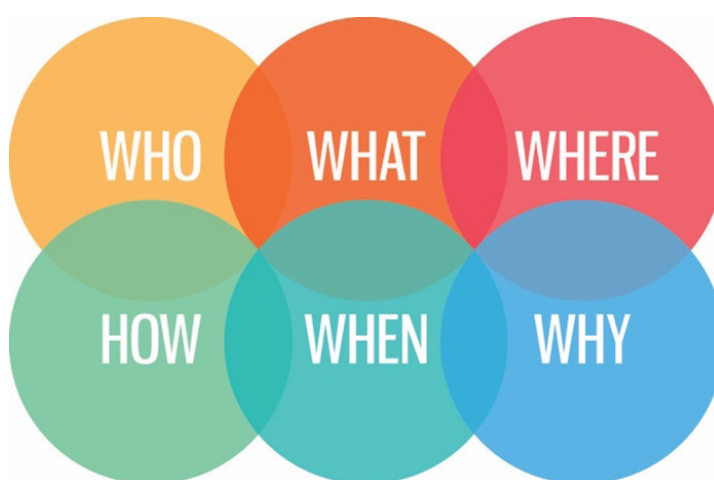
Урок 11

Проверка данных

Короткое описание урока: из этого урока вы узнаете о том, на какие вопросы нужно ответить, чтобы доверять своим данным.

Мы привыкли верить официальным документам (особенно тем, где есть цифры). Однако в данных тоже бывают ошибки, нечаянные или нарочные. В этом уроке мы будем задавать вопросы, которые помогут нам оценить источник данных и сами данные.

Чтобы проверить качество информации, вам нужно ответить на шесть классических журналистских вопросов – по-английски они звучат как 5W and 1H.



Who? Кто?

Какая организация собрала и/или опубликовала эти данные? Доверяете ли вы этой организации? Является ли эта организация надежным источником данных?

Наши данные появились из отчета «Уровень жизни населения Кыргызской Республики 2012 – 2016» и опубликованы Национальным Статистическим Комитетом КР в 2017 году.

What? Что?

Что измеряет набор данных? Адекватно ли это измерение? Понятны ли все показатели?

Перед нами таблица «Заболеваемость населения инфекционными и паразитарными болезнями». Заболеваемость измерена количеством случаев, разбитых по 21 типу заболеваний. Это не список всех заболеваний, а только инфекционных и паразитарных, значит наш материал может быть только об этих недугах. Также важно помнить, что это не все возможные инфекционные и паразитарные заболевания, а только те, что попали в эту таблицу.

Where? Где?

Где были собраны эти данные? Были ли они собраны в одном отдельно взятом регионе или во всех регионах страны?

Мы работаем с данными, собранными по всей Кыргызской Республике. Они не разбиты на регионы, и значит мы можем судить лишь о динамике по всей стране.

When? Когда?

Какой временной промежуток охватывает этот набор данных?

В таблице представлена динамика заболеваний за пять лет, с 2012 по 2016. Значит, мы можем посмотреть, насколько изменилась заболеваемость в эти годы.

Why? Почему?

С какой целью производился сбор данных?

Данные собирались в процессе регистрации заболеваний каждым врачом. Здесь важно понимать, могло ли произойти умышленное или нечаянное преувеличение или преуменьшение данных.

How? Как?

Это, пожалуй, самый важный и самый сложный вопрос.

Какова методология сбора данных? Занималась ли эта организация сбором данных самостоятельно, или она привлекла для этой цели другую компанию? Прошли ли работники специальное обучение? Эти данные реально собраны или экстраполированы из небольшой выборки? Что сказано о погрешностях и неточностях данных? Какие демографические группы могут быть не охвачены этими данными?



Попробуйте ответить на этот вопрос сами.

На заметку: если вы будете соединять и сравнивать схожие данные, удостоверьтесь, что их сбор проводился по одной и той же методологии. Также можете проверить данные, сравнив их с данными по той же теме из других источников – негосударственных или международных.

Урок 12

Организация датасета

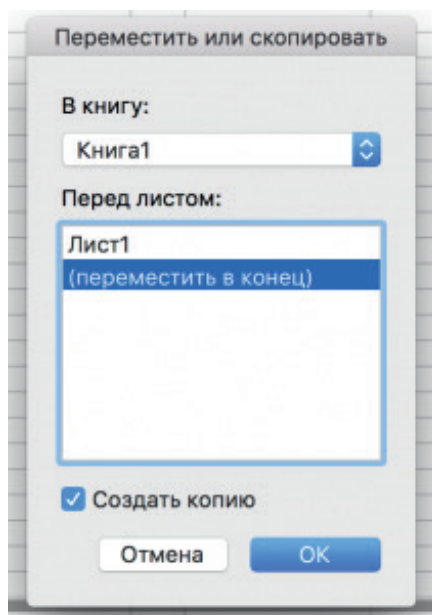
Короткое описание урока: Из этого урока вы узнаете о том, как организовать таблицу для дальнейшей работы.

Есть несколько ключевых правил при работе с данными:

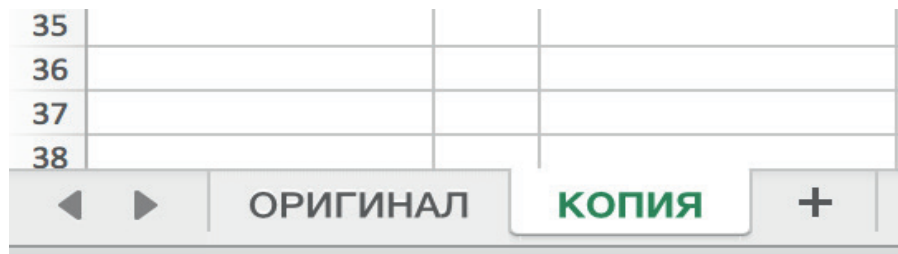
- не работать в оригинале, а только в копии
- записывать метаданные, содержащие информацию о вашем датасете, на отдельном листе

В Excel можно создать копию ваших данных следующим образом.

1. Наведя курсор на «Лист 1» и щелкнув правым щелчком мыши, выберите из меню «Переместить или скопировать», а затем «переместить в конец», «создать копию» и далее «Ок». Так вы создадите копию исходной таблицы на другом листе.



2. Двойной щелчок мыши по названию листа позволяет переименовывать его. Назовите исходник «Оригинал», а копию – «Копия».



3. Щелкнув на знак «+» в строке листов, создайте новый лист и назовите его «Мета». Переместите его в начало таблицы.



Что содержит лист с метаданными?

Основную информацию о том, какие данные содержатся у вас в таблице. Помните, что мы взяли эту таблицу из большого отчета, а работаем с ней как с отдельным файлом, при этом мы можем комбинировать этот файл с другими, посылать коллегам и выкладывать для пользования аудитории.

Более того, эта информация будет особенно полезна вам самим, когда вы вернетесь к этой таблице через некоторое время. Для этого ваш лист с метаданными должен полностью отвечать на вопросы о происхождении данных и степени доверия к ним.

Ваша программа-минимум – это следующие четыре пункта:

- **О чем ваш набор данных.** Это может быть заголовок таблицы. Но лучше, если он будет включать в себя временной и территориальный показатели. В нашем случае это «Заболеваемость населения инфекционными и паразитарными болезнями, 2012-2016, КР».
- **Публикация или база данных, которая содержит этот набор данных и ссылка на**

эту публикацию. Запишем «Уровень жизни населения Кыргызской Республики 2012-2016, Бишкек, 2017» и дадим ссылку.

- **Контакты людей, ответственных за публикацию этих данных.** Эти контакты можно найти на том же сайте Национального Статистического Комитета, в разделе «Контакты». Скопируйте контактную информацию оттуда и добавьте на лист с метаданными.

[Главная](#) / [Официальная статистика](#) / [Уровень жизни населения](#)

Уровень жизни населения



Контакты

УПРАВЛЕНИЕ СТАТИСТИКИ УСТОЙЧИВОГО РАЗВИТИЯ И ОКРУЖАЮЩЕЙ СРЕДЫ

Начальник
управления

+996 (312) 32-46-55

- **Дата скачивания вами этой таблицы.** Это важно, чтобы задокументировать дату сохранения информации. Данные могут обновляться, изменяться или вовсе исчезать с сайтов. Данные, которые вы получили по запросу об информации, могут устареть или быть опровергнуты впоследствии. Ставя дату, вы обезопасите себя от недоразумений в будущем.

	A	B	C	D	E	F	G
1	Заболееваемость населения инфекционными и паразитарными болезнями, 2012-2016, КР						
2	Уровень жизни населения Кыргызской Республики 2012-2016, Бишкек, 2017						
3							
4	http://stat.kg/ru/publications/uroven-zhizni-naseleniya-kyrgyzskoj-respubliki/						
5	Управление статистики устойчивого развития и окружающей среды						
6	+996 (312) 32-46-55						
7							
8	Дата скачивания:						
9	01.09.2018						
10							

Должно получиться примерно так:

В этот же лист вы можете помещать любые примечания, аббревиатуры, определения, а также ваши личные заметки, которые вы будете делать в ходе дальнейшей очистки и анализа данных.

Урок 13

Чистка данных

Короткое описание урока: Из этого урока вы узнаете о том, как подготовить данные к анализу.

При чистке данных необходимо получить в результате таблицу, в которой есть только один ряд заголовков, а данные в столбцах под ними соответствуют заголовкам по формату и содержанию. Это позволит вам анализировать и визуализировать данные.

Почистим нашу таблицу о заболеваемости в Кыргызстане.

Напоминаем: работайте в копии ваших данных!

Правило 1. Только один ряд заголовков.

В нашем случае, ряд заголовков – это первая строка с обозначением лет. Обратите внимание, что заголовок в первом столбце отсутствует. Допишите его в ячейке A1.

	A	B	C	D	E	F
1	Заболевания	2012	2013	2014	2015	2016
2	Брюшной тиф и паратифы	50	69	71	45	36

Правило 2. В одном столбце только один тип данных.

Если ваш столбец назван «Заболевания», то в нем могут быть только названия заболеваний.

Если у вас в столбце два типа данных, например, заболевание и область, вам надо создать еще один столбец, озаглавить его «Область» и перенести все данные об областях туда.

Обратите внимание на строки 6 и 7 в таблице: «гастроэнтериты, колиты, вызванные установленными возбудителями». Строка 7 не имеет данных, потому что название заболевания разделилось

на две строки. Чтобы исправить это, в строке 7 запишите полное название заболевания, а строку 6 удалите. Для этого наведите курсор на номер строки, нажмите правой кнопкой мыши и выберите «удалить».

	A	B	C	D	E	F
1	Заболевание	2012	2013	2014	2015	2016
2	Брюшной тиф и паратифы	50	69	71	45	36
3	Корь	-	1	308	17783	-
4	Паротит эпидемический	377	690	892	569	328
5	Вирусный гепатит (включая сывороточный)	22226	20486	12024	11252	14615
6	Острые кишечные инфекции	21665	18466	18606	17894	18398
7						
8	Вырезать ⌘X	10444	9039	9819	9680	10777
9	Копировать ⌘C	1673	1540	1809	2131	1503
10	Вставить ⌘V	2071	431	694	585	1711
11	Специальная вставка... ^⌘V	191725	142012	159934	171640	261366
12		273	199	93	81	82
13	Вставить вырезанные ячейки	1	-	-	-	1
14	Вставка	63	94	113	280	195
15		8	12	28	90	16
16	Удалить	2296	1364	1124	911	925
17	Очистить содержимое	1310	1067	898	908	750

Должно получиться вот так:

6	Острые кишечные инфекции	21665	18466	18606	17894	18398
7	Гастроэнтериты, колиты	10444	9039	9819	9680	10777
8	Бактериальная дизентерия	1673	1540	1809	2131	1503

Также обратите внимание на строку 18, «ВИЧ – инфекция 1». Действительно ли это название заболевания? В случаях сомнения всегда возвращайтесь к оригиналу, то есть к данным в отчете. Вы увидите, что 1 – это сноска, которая говорит о том, что данные по ВИЧ представлены по числу граждан Кыргызстана.

В таком случае, мы должны добавить эту информацию в лист с метаданными, а цифру «1» удалить.

Педикулез	131	82	127	562	414
ВИЧ-инфекция	704	480	612	605	714
Эхинококкоз	930	1049	1181	1134	963

Правило 3. В одном столбце только данные одного формата

Что это означает? Данные могут быть представлены в текстовом или числовом формате. Числовой формат имеет несколько подвидов: простое число, процент, календарная дата.

В каком формате должны быть данные в столбце «заболевания»? А в столбцах по годам?

Это не просто формальность. Если данные не записаны как числа, то Эксель не сможет проводить с ними различные математические операции. А вы, таким образом, не сможете узнать, растет или падает число заболеваний и в каком году было больше всего случаев.

Как определить, в каком формате записаны данные? Как правило, Эксель форматирует текст по левому краю, а числа по правому. Проверьте, так ли это у вас в таблице.

	A	B	C	D	E	F
1	Заболевания	2012	2013	2014	2015	2016
2	Брюшной тиф и паратифы	50	69	71	45	36
3	Корь	-	1	308	17783	-

Правило 4. Обратите внимание на поля, в которых проставлен дефис.

В нашем случае, дефисом обозначены отсутствующие данные. Всегда будьте очень внимательны с такими полями: означают ли они, что значение в них равно нулю, или же по этому показателю не удалось собрать данные?

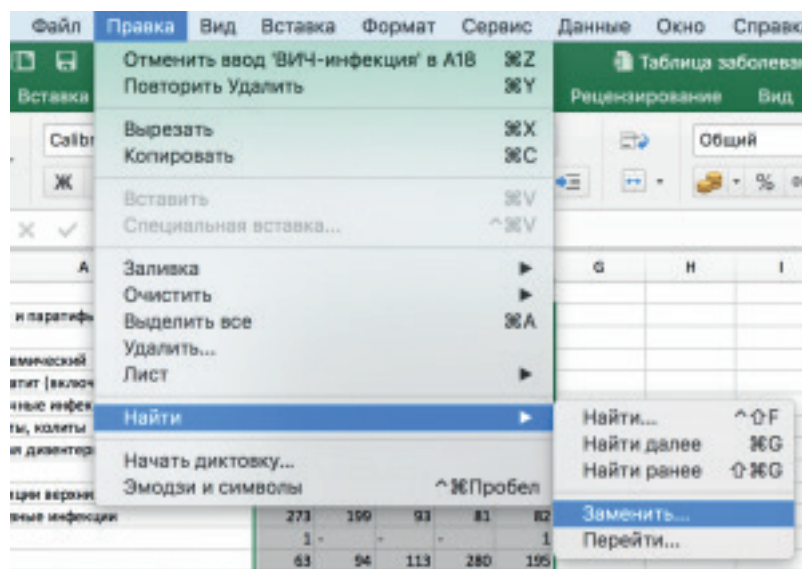
Если первое, то мы вправе заменить «-» на 0, а если второе, то мы не сможем работать с данными

этому заболеванию, либо же нам придется найти отсутствующую информацию из другого источника и добавить в таблицу.

В нашем случае «-» действительно означает 0, поэтому мы выполним простую операцию, которая автоматически заменит все дефисы на 0.

Для этого, выделите все ячейки с данными, и найдите команду «Найти и заменить».

	В	С	Д	Е	Ф
	2012	2013	2014	2015	2016
	50	69	71	45	36
-		1	308	17783	-
	377	690	892	569	328
	22226	20486	12024	11252	14615
	21665	18466	18606	17894	18398
	10444	9039	9819	9680	10777
	1673	1540	1809	2131	1503
	2071	431	694	585	1711
А	191725	142012	159934	171640	261366
	273	199	93	81	82
	1	-	-	-	1
	63	94	113	280	195
	8	12	28	90	16
	2296	1364	1124	911	925
	1310	1067	898	908	750
	131	82	127	562	414
	704	480	612	605	714
	930	1049	1181	1134	963
	3	4	-	1	6
	6	16	5	19	12
	1	1	1	2	-



В вашей версии операционной системы это может быть в другой строке меню.

Теперь, в поле «Найти» напишите «-», а в поле «Заменить» не пишите ничего.

Нажмите «Заменить все».

Заменить

Найти:

-

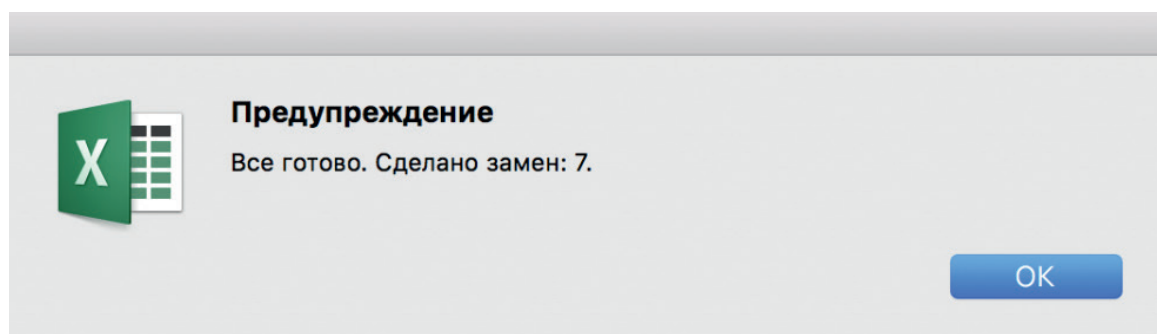
Искать: на листе Учитывать регистр

Поиск: по строкам Ячейка целиком

Заменить на:

Заменить Заменить все Заккрыть Найти далее

Готово! Мы сделали 7 замен.



У нас получилась таблица, в которой 6 столбцов, 22 строки. Она готова к анализу. Поздравляем!

Урок 14

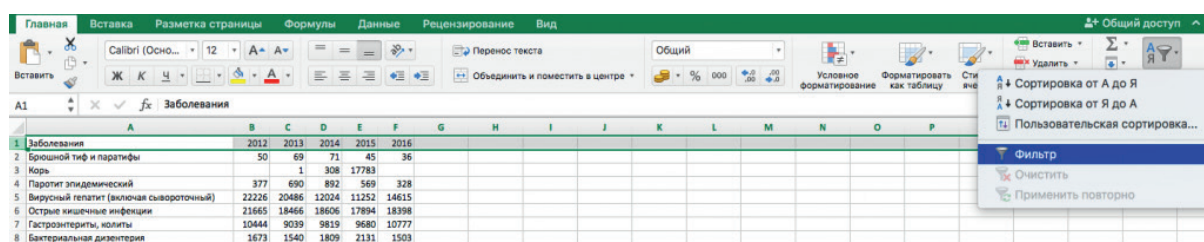
Накладываем фильтр

Короткое описание урока: В этом модуле вы узнаете, как и для чего применять функции «сортировка» и «фильтр».

Анализировать данные в таблице можно различными путями: вы можете искать среднее значение, высчитывать процентный рост и строить графики. Но самый простой метод анализа – это сортировка, которая позволит найти минимальные и максимальные значения в столбцах.

Посмотрим на столбцы нашей таблицы. У нас есть данные о заболеваемости по годам. Рассмотрим случаи заболеваний в 2016 году, просто отсортировав их. Как и раньше, работаем в программе Excel.

1. Выберите первую строку в таблице. Сделайте это, щелкнув по цифре 1, которая является номером строки.
2. Нажмите на функцию «Фильтр» в меню «Сортировка и фильтр».



3. Теперь вы можете увидеть треугольники рядом с именами столбцов в первой строке.
4. Нажмите треугольник рядом с названием заболеваний. Отсортируйте по алфавиту ("по возрастанию").

Сортировка

А↓

По возрастанию

З↓

По убыванию

По цвету:

Нет

Фильтр

По цвету:

Нет

Выберите

Поиск

☒ (Выделить все)

☒ Бактериальная дизентерия

☒ Бешенство

☒ Бруцеллез

☒ Брюшной тиф и паратифы

☒ Вирусный гепатит (включая сыворотку)

☒ ВИЧ-инфекция

☒ Гистоплазмоз

Очистить фильтр

- Вы должны увидеть список заболеваний по алфавиту. Числовые значения отсортировались вместе с текстовыми: Excel сортирует не просто ячейки в одном столбце, а все строки целиком.
- Переходим на 2016 год в первой строке и нажимаем на треугольник. Выбираем функцию сортировки по убыванию.

2	Острые инфекции верхних дыхательных путей	191725	142012	159934	171640	261366
3	Острые кишечные инфекции	21665	18466	18606	17894	18398
4	Вирусный гепатит (включая сывороточный)	22226	20486	12024	11252	14615
5	Гастроэнтериты, колиты	10444	9039	9819	9680	10777
6	Грипп	2071	431	694	585	1711
7	Бактериальная дизентерия	1673	1540	1809	2131	1503
8	Эхинококкоз	930	1049	1181	1134	963
9	Бруцеллез	2296	1364	1124	911	925
10	Чесотка	1310	1067	898	908	750
11	ВИЧ-инфекция	704	480	612	605	714
12	Педикулез	131	82	127	562	414
13	Паротит эпидемический	377	690	892	569	328
14	Коклюш	63	94	113	280	195
15	Сальмонеллезные инфекции	273	199	93	81	82
16	Брюшной тиф и паратифы	50	69	71	45	36
17	Менингококковая инфекция	8	12	28	90	16
18	Сибирская язва	6	16	5	19	12
19	Малария	3	4		1	6
20	Дифтерия	1				1
21	Бешенство	1	1	1	2	
22	Корь		1	308	17783	
23						
24						
25						

7. Можно убрать случаи с наименьшим количеством – 1, 6. Нажмите на синие квадратики и уберите галочки у этих значений.
8. Теперь вы успешно отфильтровали свой набор данных.
9. Можно пойти дальше и поиграть с фильтрами – данные не будут удалены, они просто не будут отображаться.

Задание: Какое заболевание было самым частым в 2012 году, исключая острые инфекции верхних дыхательных путей? Изменилась ли ситуация в 2016 году?

Урок 15

Базовый анализ данных, часть 1

Короткое описание урока: На этом уроке вы научитесь основным приемам анализа данных.

Итак, перед нами таблица с инфекционными и паразитарными заболеваниями с 2012 по 2016 год. Таблица как человек: мы можем задавать ей вопросы и получать ответы. Помогут в этом формулы Excel. С помощью фильтра мы уже ответили на такие вопросы, как:

- Каким инфекционным/паразитарным заболеванием заразились больше всего человек в 2012 году? А в 2016?
- Каким инфекционным/паразитарным заболеванием страдали меньше всего человек?

Задание: Придумайте еще несколько вопросов к этой таблице.

Вот несколько вопросов, которые придумали мы:

- Сколько всего человек заразились каждым заболеванием за пять лет?
- Какое заболевание самое многочисленное?
- Сколько человек в среднем болеет в год тем или иным заболеванием?
- За последний год, заразились ли этим заболеванием больше или меньше обычного?

Теперь, когда у нас есть вопросы, попытаемся ответить на некоторые из них. Для этого мы будем использовать формулы Excel. Они позволяют делать вычисления внутри ячеек таблицы:

- Все формулы в Excel начинаются со знака равно.
- Основные вычисления в Excel можно делать базовыми математическими знаками: + для суммы, – для разности, / для деления и * для умножения.

Считаем суммы, используя функцию СУММ

Например, чтобы подсчитать, сколько человек заразились брюшным тифом и паратифами за 2015 и 2016 год в сумме, нам нужно встать в пустую ячейку и написать формулу:

= E2+F2

	A	B	C	D	E	F	G
1	Заболевания	2012	2013	2014	2015	2016	
2	Брюшной тиф и паратифы	50	69	71	45	36	=E2+F2
3	Корь		1	308	17783		

E2 и F2 можно написать самому, а можно просто кликнуть на эти ячейки. Нажмите Enter, и получите результат: 81.

Как подсчитать всех, кто заразился брюшным тифом и паратифами за пять лет? Можно отдельно прописать каждую ячейку, но это будет долго. Представьте, что у нас не 5 лет, а гораздо больше. Для этого в Экселе существуют функции, и сейчас мы используем одну из них.

Функции – это команды в Экселе, которые позволяют делать математические вычисления с указанными блоками. В нашем случае, функция суммы поможет посчитать общее число заболеваний тифом за пять лет.

Для подсчета суммы существует функция «СУММ» (или SUM в английской версии). Большинство функций выражаются аббревиатурами.

Запишите в ячейке: =СУММ или =SUM

Как только вы начнете набирать функцию, она появится в выпадающем списке. Нажмите на нее.

	A	B	C	D	E	F	G
1	Заболевания	2012	2013	2014	2015	2016	
2	Брюшной тиф и паратифы	50	69	71	45	36	=СУММ()
3	Корь		1	308	17783		

Теперь в скобках надо записать диапазон ячеек, сумму которых надо подсчитать. Нажмите на первую ячейку – B2 (заболеваемость за 2012 год) и ведите, не отпуская мышку, до F2(заболеваемость за 2016 год). У вас должна получиться формула

=СУММ(B2:F2) или =SUM(B2:F2)

	A	B	C	D	E	F	G
1	Заболевания	2012	2013	2014	2015	2016	
2	Брюшной тиф и паратифы	50	69	71	45	36	=СУММ(B2:F2)
3	Корь		1	308	17783		

Нажмите Enter, получите результат: 271. 271 человек заразился брюшным тифом и паратифами в Кыргызстане с 2012 по 2016 год.

Давайте озаглавим наш столбец как «Сумма 2012-2016».

Как подсчитать теперь такую же сумму по каждому заболеванию? Можно снова ввести формулу, а можно скопировать формулу по столбцу вниз. Для этого встаньте в правый нижний угол ячейки G2. Курсор вашей мышки должен превратиться в черный тоненький крестик. Сделайте двойной щелчок, и ваш столбец заполнится до конца таблицы.

	G
016	Сумма 2012-2016
36	271
	18092
328	2856
615	80603
398	95029
777	49759
503	8656
711	5492
366	926677
82	728
1	2
195	745
16	154
925	6620
750	4933
414	1316
714	3115
963	5257
6	14
12	58
	5

Теперь мы можем наложить фильтр на первую строку и отсортировать данные в столбце G от максимального к минимальному. Получаем ответ на наш первый вопрос: за последние пять лет, топ 5 инфекционных и паразитарных заболеваний – инфекции дыхательных путей, кишечные инфекции, гепатит, гастроэнтериты и корь.

	A	B	C	D	E	F	G
1	Заболевания	20	20	20	20	20	Сумма 2012-20
2	Острые инфекции верхних дыхательных путей	191725	142012	159934	171640	261366	926677
3	Острые кишечные инфекции	21665	18466	18606	17894	18398	95029
4	Вирусный гепатит (включая сывороточный)	22226	20486	12024	11252	14615	80603
5	Гастроэнтериты, колиты	10444	9039	9819	9680	10777	49759
6	Корь		1	308	17783		18092

Сколько всего человек заразились каждым заболеванием за пять лет? Какое заболевание самое многочисленное? Получается, за 5 лет почти миллион человек – 926 677 – страдал от инфекции горла и легких. Но так ли это?

На самом деле, перед нами таблица, где зафиксированы **случаи** заболевания. Если один человек болеет из года в год, он будет из года в год попадать в эту таблицу. Более того, если он или она

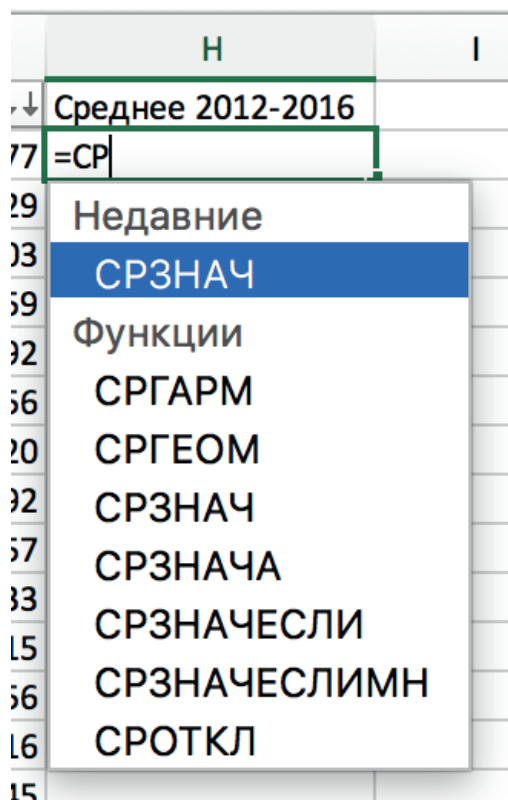
заболел два раза в одном году, то он два раза попадет в статистику по одному году. И наоборот, если кто-то лечился дома и не ходил ко врачу, его в статистике не будет.

Тем не менее, мы все равно видим, что инфекции верхних дыхательных путей – самое частое заболевание в таблице. Инфекциями желудка болеют на порядок реже.

Используем функцию СРЗНАЧ

А сколько случаев в среднем в год по каждому заболеванию? Ответив на этот вопрос, мы сможем понять общую картину заболеваемости: вместо данных за отдельно взятый год мы увидим основные тенденции. Для этого нам нужно посчитать средние значения.

Сделаем это в отдельном столбце. Озаглавьте его «Среднее 2012-2016». Начните писать формулу: =СР в ячейке Н2, и выберите формулу СРЗНАЧ (AVERAGE) из выпадающего списка.



Вам нужен тот же диапазон, что и при вычислении суммы. Ваша формула должна выглядеть так: =СРЗНАЧ(B2:F2) или =AVERAGE(B2:F2).

Нажмите Enter, получите результат: в среднем, за год регистрируется 185 тысяч инфекций дыхательных путей. Скопируйте формулу вниз по столбцу двойным щелчком по ячейке H2.

Как мы посчитали это значение? Формула СРЗНАЧ лишь берет сумму и делит её на число ячеек. Вы можете убедиться в этом сами: поделите число дыхательных инфекций за все годы из столбца «Сумма» на 5.

Обратите внимание, что ваши значения имеют знаки после запятой. Это потому, что при делении не всегда получаются целые числа. Тем не менее, вы можете округлить значения. Выделите их и сократите знаки после запятой кнопкой «уменьшить разрядность».

	Н	И
Среднее 2012-2016	185335	
2017	19006	
2019	16121	
2020	9952	
2021	6031	
2022	1731	
2023	1324	
2024	1098	
2025	1051	
2026	987	
2027	623	
2028	571	
2029	263	
2030	149	
2031	146	
2032	54	
2033	31	
2034	12	
2035	4	
2036	1	
2037	1	

Теперь вы можете сравнить, например, показатели за 2016 год со средним значением. Это даст вам ощущение того, велик или мал показатель за последний год. Например, в 2016 году с инфекциями верхних дыхательных путей было зарегистрировано 261 366 случаев. В среднем за год, если брать последние 5 лет, регистрируются 185 тысяч. О чем вам это говорит?

На заметку: теперь у вас есть вся информация, чтобы ответить на вопросы из начала этого урока.

Урок 16

Базовый анализ данных, часть 2

Короткое описание урока: это продолжение основных приемов анализа данных. В этом уроке мы научимся считать процентный рост.

В предыдущем уроке мы посчитали сумму и среднее значение заболеваний за 5 лет. Что нам еще интересно узнать из этой таблицы?

- В каком году были вспышки заболеваний?
- По каким заболеваниям число случаев росло, а по каким – падало за эти пять лет?
- Какое заболевание больше всего выросло в количестве?

Вспышки заболеваний – это отклонения от нормы. Чтобы увидеть их, нам нужно сначала рассчитать «норму», то есть показатель, который не удивляет. Мы уже частично сделали это в предыдущем уроке: среднее значение за 5 лет – это и есть своеобразная норма, и если ваш показатель больше ее, то этим заболеванием болеют больше, чем обычно.

Как увидеть, что показатель за 2016 год больше среднего? Можно просто посчитать разницу:

Для острых инфекций верхних дыхательных путей это 261 366 – 185 335, то есть на 76 тысяч больше. Но подсчет разницы не даст нам возможности сравнивать заболевания между собой. Например, показатели гриппа выросли всего на 613 случаев в абсолютном значении, но это 1711 случай в 2016 по сравнению с 1098 в среднем за каждый год. Чувствуете разницу?

Как можно эту разницу выразить численно?

Для этого можно посчитать, во сколько раз показатели за 2016 год больше или меньше среднего. Тогда мы сможем сравнивать результаты по разным заболеваниям между собой.

Для этого нам нужно поделить показатель за 2016 год на среднее значение. Запишите следующую формулу в ячейке I2: $=F2/H2$

Озаглавьте столбец как «2016/СРЗНАЧ». Скопируйте формулу вниз по столбцу, наложите фильтр, отсортируйте значения от максимального к минимальному и посмотрите, какие заболевания в 2016 больше всего отличались от нормы.

Разделив значения за 2016 год на норму, мы теперь можем сказать, что в этом году инфекциями легких и горла болели в 1.4 раза чаще обычного. Видите, получается совсем другая история.

	A	B	C	D	E	F	G	H	I
1	Заболевания	2015	2014	2013	2012	2011	Сумма 2012-2015	Среднее 2012-2015	2016/СРЗНАЧ
2	Малярия	3	4		1	6	14	4	1,7
3	Педикулез	131	82	127	562	414	1316	263	1,6
4	Грипп	2071	431	694	585	1711	5492	1098	1,6
5	Острые инфекции верхних дыхательных путей	191725	142012	159934	171640	261366	926677	185335	1,4
6	Коклюш	63	94	113	280	195	745	149	1,3

Задание: посчитайте по аналогичной формуле показатели за 2015 год, сравнив их со средним значением, и найдите историю. О чем она?

Теперь давайте посчитаем, по каким заболеваниям число случаев росло, а по каким – падало за эти пять лет?

Для этого разберем формулы темпа роста и темпа прироста. Темп роста – это отношение текущей величины к базовой. Он выражается в процентах и не может быть меньше нуля. Например, темп роста в 100% означает, что показатель не изменился.

Запишем в ячейке J2 формулу темпа роста для малярии. За базовый год возьмем 2012. Формула будет выглядеть как:

=F2/B2

Озаглавьте столбец как «темп роста» и скопируйте формулу вниз по столбцу. Переведите показатели в процентный формат, кликнув на значок «%» на главной вкладке меню.

Процентный % 000 ←,0 ,00 →,00 Условное форматирование		
Н	I	J
Среднее 2012-2015	2016/СРЗН	Темп рос
4	1,7	200%
263	1,6	316%
1098	1,6	83%
185335	1,4	136%
149	1,3	310%
623	1,1	101%
9952	1,1	103%
12	1,0	200%
1	1,0	100%
19006	1,0	85%
1051	0,9	104%
16121	0,9	66%
1731	0,9	90%
987	0,8	57%
1324	0,7	40%
54	0,7	72%
571	0,6	87%
146	0,6	30%
31	0,5	200%
6031	0,0	#ДЕЛ/0!
1	0,0	0%

Отсортируйте данные от максимального к минимальному. Что получилось? Корь выпала из анализа, потому что в 2012 году было зарегистрировано 0 случаев кори. Педикулез и коклюш – заболевания, показатели по которым выросли более чем в 3 раза, то есть на 316 и 310 процентов соответственно.

Посмотрите на конец списка. Темп роста для сальмонеллезных инфекций – 30%. Увеличилось или

уменьшилось значение по этому заболеванию за 5 лет?

Получается, значения меньше 100% сигнализируют о том, что показатель не вырос, а уменьшился. Для того, чтобы сразу подсчитать, на сколько уменьшился показатель, используется формула темпа прироста. Фактически это разница между тем, что стало и тем, что было. Высчитывается формула как "Темп Роста – 1". Темп прироста тоже выражается в %, но далеко не всегда больше 0. Темп прироста в 100% означает, что показатель вырос в 2 раза.

Запишем формулу в ячейке K3 для педикулеза:

=F3/B3-1

Скопируем формулу вниз по столбцу, переведем в формат %, озаглавим столбец и отсортируем данные на этот раз от минимального к максимальному. Обратите внимание на болезни, темп прироста по которым меньше 0. Это означает, что там становилось меньше случаев заболевания. Вот еще одна история для вашего материала.

	A	B	C	D	E	F	G	H	I	J	K
1	Заболевания	2012	2013	2014	2015	2016	Сумма 2012-2016	Среднее 2012-2016	2016/СРЭН	Темп роста	Темп прироста
2	Бешенство	1	1	1	2		5	1	0,0	0%	-100%
3	Сальмонеллезные инфекции	273	199	93	81	82	728	146	0,6	30%	-70%
4	Бруцеллез	2296	1364	1124	911	925	6620	1324	0,7	40%	-60%
5	Чесотка	1310	1067	898	908	750	4933	987	0,8	57%	-43%
6	Вирусный гепатит (включая сывороточный)	22226	20486	12024	11252	14615	80603	16121	0,9	66%	-34%
7	Брюшной тиф и паратифы	50	69	71	45	36	271	54	0,7	72%	-28%

Урок 17

Этические аспекты в области журналистики данных

Короткое описание урока: Любая журналистика данных требует данных, но нахождение или сбор данных создает множество логистических и этических проблем.

Данные являются прекрасным источником для вас как для журналистов, поскольку они придают достоверность вашим источникам и могут помочь наглядно объяснить общественности сложные темы. Данные также являются подходящим инструментом для получения более глубокого понимания событий, для выявления связей, которые не так уж очевидны для широкой общественности, и для «выяснения истины», что считается одним из важнейших этических принципов работы журналистов.¹

Серьёзность журналистики данных связана с достоверностью данных. А в некоторых случаях даже официальные данные могут быть неточными. Но даже если данные очень точны, журналист может задавать неактуальные вопросы, анализировать неверные данные или упускать из виду важные материалы в своих публикациях. Другие проблемы могут возникнуть в случае отсутствия данных, недостаточности методов анализа, необоснованности выводов или неполноты общей картины по данному вопросу. По этой причине анализ результатов требует точности и тщательного изучения, чтобы избежать предвзятости и предотвратить некритичное восприятие журналистами того, что они надеются увидеть в данных. Кроме того, вам как журналистам необходимо убедиться в том, что логика интерпретации ваших собственных выводов верна.

В общем, этические соображения в информационной журналистике существенно не отличаются от этических соображений в любой другой области журналистики. В принципе журналистика данных сталкивается с теми же этическими проблемами, что и традиционная журналистика, например, в отношении неправдивого изложения материала или угрозы неприкосновенности частной жизни.

¹ Содержащиеся и рекомендации в этой главе взяты из различных публикаций экспертов в области журналистики данных и адаптированы к ним, в том числе Bradshaw, Paul (2013): Ethics in data journalism: accuracy. Online journalism blog. (<https://onlinejournalismblog.com/2013/09/13/ethics-in-data-journalism-accuracy/>); McBride, Rebekah E.D. (2016): The Ethics of Data Journalism. (<http://digitalcommons.unl.edu/journalismprojects/9/>); Kuutti, Heikki (2016): Ethics of data journalism. Four ethical phases in the working process. (<https://jyx.jyu.fi/handle/123456789/58616>). Gray, Jonathan & Chambers, Lucy & Bounegru, Liliana (eds.) (2012): The Data journalism handbook. European Journalism Center. (<https://datajournalismhandbook.org/>); LaFleur, Jennifer (2014): Data journalism tips from data driven ethics. (<https://www.scu.edu/ethics/focus-areas/journalism-ethics/resources/data-journalism-tips-/>).

Однако этические проблемы в изложении данных трудно определить, в отличие от неправильно написанных цитат собеседника или явно ложной информации в материале.

Более того, хотя журналистика данных в значительной степени является частью этики и практики традиционных журналистских информационных технологий, используемых в этом виде журналистской практики, она породила дополнительные опасения относительно доступа к материалам данных, использования данных для анализа, прозрачности в публикации результатов и открытости для получения отзывов общественности и исправления ошибок в материалах.

Этические проблемы журналистики данных связаны с различными фазами журналистского рабочего процесса и, как описано ниже, каждая фаза требует особых этических соображений.

Проблемы поиска и сбора данных

Любая журналистика данных требует наличие данных, но нахождение или сбор данных создаёт множество логистических и этических проблем. Несмотря на трудности, важно помнить, что этот шаг является основой истории, поэтому понимание технических потребностей и этических проблем имеет решающее значение для создания качественного конечного продукта. С точки зрения сбора данных, необходимо будет принять несколько решений на этом пути, но одним из первых решений, которое необходимо будет принять, будете ли вы и ваши коллеги собирать данные самостоятельно или использовать уже существующую базу данных.

Если вы используете уже существующие базы данных, например государственные или корпоративные и коммерческие документы, или результаты обследований или политических опросов, вам следует рассмотреть, откуда поступают данные, кто их собирает, каковы были мотивы данного лица, в каком контексте данные предоставляются и т.д.

При изучении наборов данных вы, как журналист, должны задавать те же вопросы, которые задаёте в отношении любого другого источника. Первый вопрос, который вам необходимо задать себе по поводу любых данных, - в чем заключается личная заинтересованность лица, предоставляющего данные? Какой вопрос являлся главным для исследования? Помните, что как бы нам ни хотелось верить, что опросы и исследования проводятся исключительно из любопытства, люди обычно относятся к ним с предсказанием результата или предположением. Особенно при представлении результатов обследований следует проявлять «соответствующий скептицизм» и, при необходимости, включать в них описание методологии и методов составления выборки, включая число респондентов.

При работе с опросами всегда запрашивайте доступ к исходным данным, включая все заданные вопросы.

В ходе освещения политических выборов или других показателей, таких как уровень безработицы, особенно важно сообщать о допустимой погрешности. Это диапазон, в котором скорее всего будут находиться «истинные» показатели. Когда вы информируете об опросах, вы должны выявить организацию, которая проводила опрос, вопросы, метод и размер выборки, а также более широкую тенденцию во всех опросах или в конкретном опроснике. Периодически вы захотите составить отчёт на основе статистики. Иногда такая статистика предоставляется всем информационным агентствам. В таких случаях важно иметь собственные источники для перепроверки. Не обращайтесь к тем, которыми пользуются другие. Они могут быть ненадёжными. Создайте свою собственную сеть доверенных контактов и обратитесь к ним.

Другими важными вопросами при изучении наборов данных являются: Имеете ли вы все данные, необходимые для вашей истории, или же их может быть не хватает? Как была собрана эта информация - помните, что неприкосновенность частной жизни находится под угрозой в цифровом мире. Как были сформулированы вопросы, например обследования, и какие вопросы были использованы для составления данных? Насколько обоснованным и достаточным является метод анализа, насколько обоснованными являются выводы, является ли общая картина достаточно широкой и точен ли контекст? Есть ли второй независимый источник той же информации или иной интерпретации?

Как уже упоминалось выше, когда-нибудь вам может понадобиться или вы захотите собрать данные самостоятельно. При самостоятельном сборе данных, например, с помощью онлайн-опросов или краудсорсинга (процесс привлечения общественности к сбору данных), целесообразно использовать некий вид «практического доверия» в сочетании с «здравым скептицизмом». Например, если вы хотите собрать данные о коммунальных услугах, вы можете попросить пользователей также отправить вам копию их счетов. Это может ввести в неловкое положение некоторых пользователей, но повысит доверие к данным, полученным с помощью краудсорсинга.

Вы должны также знать, что, хотя технология дала нам множество вариантов голосования, не все они созданы равными. Каждый метод имеет свои плюсы и минусы. Тип используемого вами опроса будет зависеть от того, кого вы надеетесь охватить и насколько широк диапазон мнений, с которыми вы надеетесь столкнуться.

Журналисты данных также часто принимают участие - и организывают - «дни взлома» или «ха-

катоны», направленные на открытие и объединение данных и создание приложений, или работают с внешними агентствами для анализа данных, собранных любой из сторон. Это может привести к конфликту между культурами из-за различных этических принципов. В этих случаях важно знать этические представления различных сторон и чётко понимать роль журналиста.

Однако доступ к данным - это только половина проблемы, так как при работе с ними вам также необходимо уметь собирать, интерпретировать и организовывать собранные данные.

Интерпретация и анализ данных

Доступ к данным может привести журналистов к поспешному и небрежному использованию данных. Однако при работе над историей данных важно помнить, что данные являются лишь объектом анализа и требуют дополнительных источников для завершения этой истории. Главной ценностью журналистики данных, как и всех других направлений журналистики, является точность. Поэтому вы должны стремиться излагать факты точно и правдиво, создавать надлежащий контекст для представляемых вами историй и не позволять себе видеть только то, что вы надеетесь увидеть. Данные не должны искусственно высказывать больше, чем они могут. Также вам следует спросить себя, высказывают ли данные именно то, что они должны делать; проводится ли анализ достаточно всесторонне и в контексте реальности.

Используя данные, вы должны быть очень осторожны, потому что неправильное использование статистики может произойти быстро. Другая этическая проблема при анализе данных связана с неправильным толкованием результатов или ложными сравнениями, которые делают журналисты. Это происходит, когда журналисты пишут свои репортажи исключительно на основе данных и без использования других независимых источников и без обращения к внешним экспертам с просьбой прокомментировать результаты. Так что создайте свою собственную сеть доверенных контактов и обращайтесь к ним.

Независимо от того, насколько захватывающе или интригующе звучат имеющиеся у вас данные, вы должны проверить их, найти возможные альтернативные результаты и спросить, что ещё могли бы объяснить ваши выводы.

Цифры, статистики, диаграммы и карты обладают неким авторитетом, которого часто не хватает другим типам информации, тем не менее ими также можно манипулировать. Каждый алгоритм, как бы он ни был написан, содержит человеческое, а значит и редакционное, суждение. А в некото-

рых случаях даже официальные данные могут быть неточными. Как журналист, вы должны быть осторожны как в отношении доверия к числовым и графическим источникам, так и в том, как вы представляете свои собственные истории с помощью цифр и графических средств.

И даже решения, которые вы принимаете о том, какие данные включать и исключать, добавляют определённый уровень перспектив к предоставляемой информации. Таким образом, чтобы обеспечить ответственную публикацию данных, вы должны спросить себя о необходимости опубликования чего-либо, причинах, по которым вы не публикуете данные, и о том, как лучше всего это сделать.

В целом, неверной является гипотеза о том, что данные всегда достоверны или компьютерный анализ не допускает ошибок. Журналисты должны относиться к данным с таким же скептицизмом как и к человеческим источникам. И если вы обнаружите ошибки в данных, то подчеркнуть эти неточности является вашим этическим долгом. Более того, если ошибки окажутся серьёзными, вы должны отказаться от публикации этих данных. Проверка данных перед публикацией это важный шаг для обеспечения точности.

Иногда, однако, размер или характер данных делает невозможной проверку каждой строки, но все же существует этический императив проверять все факты перед публикацией или требовать наличия двух источников по каждому факту.

Опубликование данных

После того, как вы получите все данные, вы должны подумать о том, как вы их представите и опубликуете. Данные, используемые в материале, должны иметь чёткую связь с результатами. А поскольку данные могут быть ошибочными, вводящими в заблуждение, вредными, смущающими или инвазивными, добавление контекста является жизненно важной частью процесса журналистики данных. Даже если данные точны, их контекст может оказаться недостаточным, а также они могут не соответствовать действительности, которую они должны описать. Поэтому необходимо иметь соответствующую справочную информацию и какие-либо руководства по интерпретации информации.

Несомненно, что абсолютные цифры и результаты измерений как таковые ничего не говорят о том, выше ли эти показатели или ниже того, чем должны быть, повышаются ли они или понижаются, являются ли они лучшими или худшими в регионе, стране или мире. Для этого абсолютные цифры должны рассматриваться в контексте численности местного населения, исторических особенностей

и даже различий в демографических характеристиках. Приведение цифр в контекст позволяет читателям лучше понять смысл данных, а вы, как журналист, с меньшей вероятностью намекаете на ложные корреляции или приводите читателей к неверным выводам.

Кроме того, визуализация данных должна быть не менее чёткой: например, диаграммы и таблицы должны, как правило, иметь нулевой уровень, чтобы избежать искажённого представления изменений как более серьёзных, чем они есть. Временные рамки следует выбирать для представления долгосрочных тенденций, а не искажать путём их отображения на этапе запуска или выбора с самого низкого или высокого за всю историю значения.

И как вы представляете результаты опросов в процентах? Будете ли вы округлять с точностью до ближайшего процентного пункта? Округление данных представляет собой большую проблему, и оно потенциально может существенно изменить цифры. Поэтому результаты обследований в процентах должны использоваться только с осторожностью и с учётом контекста.

При использовании материалов третьих сторон всегда важно чётко указывать, какие материалы были предоставлены другими лицами. Выражайтесь ясно и понятно, следите за своей речью. Скажите «в соответствии с...» или «о нем сообщают...» или «опросы предполагают» и «указывают», но никогда не говорите «доказывают» или «показывают».

Как и в традиционной журналистике, редакционные суждения в журналистике данных находятся между общественным интересом и личной жизнью. Как упоминалось выше, неприкосновенность частной жизни находится под угрозой исчезновения в нашем цифровом мире. А вторжение в частную жизнь является одной из главных этических проблем при публикации истории данных. Данные могут быть конфиденциальными или связаны не с тем человеком. Данные в составе более крупных массивов данных могут содержать дополнительную информацию, которая фактически не является необходимой для конкретной статьи. И даже анонимизированные данные могут раскрыть личную информацию.

Кроме того, очень важно при использовании данных учитывать потенциальные ограничения данных в вашей истории. Поэтому вы должны решить, как представить данные таким образом, чтобы извлечь максимальную пользу и свести к минимуму вред. Речь идёт о том, какие аспекты данных действительно важны, и что вы, как журналист, можете ограничить или отредактировать. Однако в каждом конкретном случае следует отдельно рассмотреть вопрос о том, следует ли публиковать данные и каким образом.

Очень важно помнить, что вам также может понадобиться быть осторожным в отношении защиты источников информации при публикации утечек данных. Помните, что метаданные, хранящиеся в файлах, содержат дату и место доступа, используемые компьютеры и учётные записи, а также другие данные, которые могут быть использованы для идентификации источника. Даже фотокопии или печатные материалы могут содержать невидимые цифровые водяные знаки, которые описывают, какие машины использовались для их производства и когда.

Поэтому следует избегать публикации оригинальных исходных материалов в онлайн-режиме, если подлинность источника неизвестна. По той же причине, как и журналистам, следует избегать предоставления таких материалов органам государственной власти и проявлять осторожность при хранении таких файлов онлайн в «облачных» сервисах, которые могут быть перехвачены службами безопасности или подвергнуты юридическому давлению.

В заключение, когда ваша история будет завершена, вам нужно вернуться и проверить факты по порядку. Связь между фактами, данными и документами должна по мере возможности прослеживаться. Вы должны открыто обсуждать выводы по результатам с коллегами. И для надёжности полезно найти независимых экспертов для анализа результатов.

Публикация вашей истории данных

Всегда помните, что убедительная история данных должна содержать описание использованных данных и включать происхождение, содержание и характер материала, а также методы анализа, использованные в истории.

Чтобы убедить аудиторию в достоверности как результатов анализа, так и истории в целом, необходимо сделать свою деятельность максимально прозрачной, описав данные и методы анализа. Кроме того, вам следует активно привлекать общественность к тому, чтобы она указывала на возможное отсутствие в материалах данных или ошибки при сборе или интерпретации данных, и быть готовым внести необходимые исправления в материалы.

