00:00

Лектор Амрапали Завери говорит по-английски.

00:05

Сегодня я буду говорить о вводных вещах в связанные данные.

00:14

по-английски.

00:20

Я буду говорить о жизненном цикле связанных данных и о принципах, которые стоял за этим.

00:27

по-английски.

00:30

Зачем использовать связанные данные, а также я буду приводить примеры.

00:35

по-английски.

00:39

И в завершении я поговорю о достижениях и о проблемах.

00:43

по-английски.

00:48

Во второй части презентации я поговорю о качестве данных, о направлениях и способах измерения качества актуальности связанных данных.

01:05

по-английски.

01:10

Также покажу инструменты для оценки качества связанных данных.

01:18

по-английски.

01:23

Начнем с принципов связанных данных и поговорим о том, что такое связанные данные.

01:28

по-английски.

01:36

Термин «связанные данные» относится к набору лучших практик для публикации и связывания структурированных данных в сети.

01:51

по-английски.

01:53

Существует пять принципов связанных данных.

01:57

по-английски.

02:01

Первое, это использовать URI как название для объектов.

02:08

по-английски.

02:12

URI – это универсальный идентификатор ресурсов.

02:18

по-английски.

02:25

Второй принцип – использовать URI http так, чтобы люди могли находить, искать эти имена.

02:35

по-английски.

02:27

Третий принцип состоит в том, что когда кто-то ищет по URI, должна поступать полезная информация, в которой используются стандарты РДФ спаркл.

03:06

по-английски.

03:13

И последний принцип, это принцип в том, чтобы включать ссылки на другие URI, чтобы можно было находить больше объектов.

03:27

по-английски.

03:33

Это облако связанных открытых данных на момент 2011 года.

03:38

Анг.

03:45

Все эти круги отображают, изображают наборы данных, которые были преобразованы, связаны в соответствии с этими принципами.

03:56

Анг.

04:04

Разные цвета соответствуют разным доменам.

04:07

Анг.

04:13

География. Областям знания, не доменам, областям знания. География, науки о жизни, правительство.

04:26

Англ.

04:30

Давайте посмотрим на принципы немного в больших подробностях.

04:35

Англ.

04:46

Ссылки URI должны отображать, должны идентифицировать не только документы веб и цифровой контент, но также и объекты реального мира и абстрактные понятия.

05:06

Англ.

05:17

Например, уникальные ресурсы, которые идентифицируют людей, места.

05:22

Англ.

05:27

Ну и пример абстрактных вещей, например, тип отношений, когда кто-то кого-то знает.

05:35

Анг.

05:39

Например, Иван знаем Амрапали.

05:42

Англ.

05:47

И второй принцип использования URI http.

05:51

Анг.

05:54

Который позволяет повторно использовать веб архитектуру.

05:59

Англ.

06:04

И делает акцент на веб и семантик веб.

06:11

Англ.

06:15

Потому что это позволяет дереферисинг ресурсов.

06:23

Англ.

06:29

А также повторное использование стандартных инструментов для обеспечения безопасности и так далее.

06:37

Англ.

06:40

Позже я покажу примеры этого.

06:45

Англ.

06:54

Третий принцип состоит в том, что люди и машины должны, что у людей и машин должна быть возможность получать должное представление о ресурсах.

07:05

Англ.

07:13

Для людей это должен быть html, для машин это должен быть RDF.

07:17

Англ.

07:25

Для этого существует механизм http, который называется “content negotiation”.

07:33

Англ.

07:49

Основная идея в том, что клиент http посылает http-заголовки с каждым запросом для того, чтобы показать какого типа документы предпочтительны.

08:02

Англ.

08:09

Это могут быть документы html, RDF, триплеты.

08:17

Англ.

08:25

То есть они проанализируют этот заголовок и выдадут пользователю соответствующий ответ.

08:34

Англ.

08:44

Существует две стратегии. Один вот 303 URI, другой хэш URI, но мы не будем углубляться в подробности.

08:52

Англ.

09:01

Только скажу, что оба обеспечивают тот момент, что объекты и документы, которые они описывают, не перемешиваются.

09:13

Англ.

09:18

И люди, и машины получают соответствующие представления.

09:27

Англ.

09:35

Мы переходим к четвертому принципу, который касается связывания двух наборов данных.

09:41

Англ.

09:54

Базовое определение связи, или сслыки, линк, в том, что триплет соединяет URI разных дата-сетов, наборов данных, вот это, собственно, и есть связь, линк.

10:18

Англ.

10:34

Собственно, основная идея связанных данных состоит в том, чтобы применить общую гиперссылочную архитектуру воркволдвеб к тому, чтобы распределять структурированные данные по масштабу всего мира.

11:00

Англ.

11:11

То есть в общем интернете у нас есть связи между http-документами, но и мы хотим, соответственно, соединить данные внутри сферы данных.

11:27

Англ.

11:33

И хотя уже есть набор данных, которые связаны.

11:38

Англ.

11:45

Следовательно, исследовательская задача состоит в том, чтобы создать эффективные связи с высокой точностью.

11:58

Англ.

12:00

Давайте посмотрим на один пример.

12:05

Англ.

12:10

Дипипедия – это набор данных, конвертированных из википедии.

12:19

Англ.

12:23

И это URI-ресурс Санкт-Петербург.

12:27

Англ.

12:32

Это соответствует всем принципам связанных данных.

12:36

Англ.

12:40

Чуть-чуть позже посмотрим на подробности.

12:48

Англ.

12:55

Теперь. Когда мы поговорили о принципах, давайте поговорим о том, зачем использовать связанные данные.

13:01

Англ.

13:15

Давайте посмотрим на такой пример. Допустим, мы хотим найти в сети следующие вещи: жилье поблизости к двуязычному немецко-русскому детскому сады в Лейпциге.

13:30

Англ.

13:37

Ну, или, например, исследователи, которые работают с мультимедийными темами в Восточной Европе.

13:46

Англ.

13:57

Если мы будем искать эту информацию просто в сети, то она там есть, но она скрыта.

14:06

Англ.

14:13

Например, вот база данных, в которой содержится вся информация о детских садах в Лейпциге.

14:20

Англ.

14:26

В этой базе данных содержится все о риэлтерских предложениях в Германии.

14:33

Англ.

14:39

Тогда, чтобы ответить на первый вопрос, нужна такая поисковая система, которая объединяет обе эти базы данных.

14:48

Анг.

14:51

Как это сделать.

14:55

Англ.

15:02

Нам нужно следовать 5-звездочному. 5-звездочной схеме Тима Берносали.

15:11

Анг.

15:15

Для открытой сети данных.

15:19

Англ.

15:25

Он предложил 5-звездочную модель.

15:27

Англ.

15:33

В которой первая звезда, это значит, что данные доступны в сети по открытой лицензии.

15:40

Англ.

15:44

Две звезды, это значит, что данные не просто доступны, но они и структурированы.

15:49

Англ.

15:53

Три звезды означают, что используются не проприетарные форматы.

15:56

Англ.

16:00

Четыре звезды значат, что для идентификации вещей используется URI.

16:05

Англ.

16:05

Англ.

16:12

И пять звезд, это значит, что данные связываются с данными других людей, и таким образом получается контекст.

16:20

Англ.

16:24

На следующих слайдах мы посмотрим на примеры каждого, из этих случаев.

16:29

Англ.

16:42

Вообще звезд нет, да, у этого дата-каталога, в котором есть хорошие мета-данные, и его можно найти.

17:01

Англ.

17:03

Но это только первый шаг до этого – ноль звезд.

17:09

Англ.

17:19

Соответственно, первая звезда – это сделать лицензию открытой.

17:24

Англ.

17:30

Например, публикаторы могут распространять данные, чтобы сделать их более заметными. Издатели в этом случае, да, издатели.

17:46

Англ.

17:52

Например, издательский центр «Nature» сделал свои данные доступными в качестве связанных.

18:04

Англ.

18:09

Также открытыми для общественности делает данные государство.

18:21

Англ.

18:30

Это делается в целях рекламы для увеличения трафика.

18:34

Англ.

18:42

Здесь, например, данные по фотоаппарату «Никон».

18:50

Англ.

18:56

Бывают целые сообщества, в которых создаются, в которых наборы данных создаются общими усилиями.

19:04

Англ.

19:09

Одно из таких сообществ – это “Music Brains”, в котором есть данные по открытой лицензии.

19:17

Англ.

19:24

Которые содержат информацию о музыке, записях, песнях, исполнителях и так далее.

19:33

Англ.

19:37

Две звезды – предоставить структурированные данные.

19:42

Англ.

19:46

Разрешить повторное использование.

19:49

Англ.

19:54

Это позволяет предоставлять информацию конечным пользователям в разных форматах.

19:59

Англ.

20:02

Объединение данных с другими наборами данных.

20:06

Англ.

20:09

А также анализ данных с третьей стороны.

20:16

Англ.

20:23

Есть уже известные форматы.

20:26

Англ.

20:35

Существуют форматы, которые легко повторно использовать, и которые хорошо структурированы. Это, например, exel, csv, xml.

20:46

Англ.

20:53

Не очень хорошие для повторного использования, например, просто веб-сайты, или microsoft word.

21:03

Англ.

21:09

Плох для повторного использования формат pdf, потому что из него трудно извлекать данные.

21:16

Англ.

21:22

И очень плохи для повторного использования это графики и карты без числовых значений.

21:34

Англ.

21:39

Три звезды – это предоставление данных в неприприетакрных форматах.

21:44

Англ.

21:49

Потому что инструменты, которыми пользуются специалисты, очень часто нуждаются в специальных форматах.

22:00

Англ.

22:08

И инструменты, которые могут читать эти специфические форматы, мало у кого есть, они дорогие и, соответственно, такие данные трудно использовать повторно.

22:21

Англ.

22:26

Например. Инструменты для гео-данных, статистический набор данных и так далее.

22:37

Англ.

22:44

Непроприетарные форматы это, например, csv, xml, джейсон.

22:50

Англ.

22:54

Ну, и конечно, IDF, который также может получить 4 или 5 звезд.

23:03

Англ.

23:09

4 звезды, как мы уже говорили, это когда в качестве идентификаторов используется URI.

23:15

Англ.

23:22

Например, вот URF для этого набора данных на ББС по Мадонне.

23:30

Англ.

23:38

Дизайн URI желательно должен быть постоянным и не зависеть от, и он должен быть независимым от реализации.

23:55

Англ.

24:02

Да, зачем превращать локальный идентификатор в URI?

29:09

Англ.

29:12

Чтобы сделать их уникальными на мировом уровне.

24:15

Англ.

24:19

Чтобы показать, из какого источника они происходят.

24:23

Англ.

24:25

Сделать их разворачиваемыми.

24:27

Англ.

24:49

Когда кто-то ищет этот URI.

24:53

Англ.

24:55

Что он предоставляет ценную, какую-то полезную информацию.

25:01

Англ.

25:03

И сделать их связанными, связуемыми.

25:13

Англ.

25:17

Теперь мы переходим к 5-звездочному формату.

25:19

Англ.

25:25

Уровень, на котором рекомендуется предоставлять ссылки на другие наборы данных.

25:30

Англ.

25:38

Как мы знаем, нынешняя сеть содержит гипер-ссылки с одной страницы на другую.

25:45

Англ.

26:04

Соответственно, сеть данных устроена точно также. То есть в одном наборе данных, соответственно, каждое значение будет иметь ссылку на другой набор данных, если есть какая-то связь.

26:21

Англ.

26:42

Есть. Соответственно, данные некоей компании.

26:44

Англ.

26:49

Да, и соответственно, там есть некоторые географические названия.

26:54

Англ.

27:00

Тут два набора данных содержат информацию об одном и том же городе.

27:04

Англ.

27:06

На этом основании они могут быть связаны между собой.

27:13

Англ.

27:22

Подходим к завершению первой части этой презентации.

27:25

Англ.

27:29

Подводя итоги, мы посмотрели на принципы связанных данных.

27:34

Англ.

27:36

Использование URI для обозначения объектов.

27:40

Англ.

27:44

Предоставление URI http.

27:48

Англ.

27:54

И когда кто-то смотрит URI, должна предоставляться полезная информация.

27:59

Англ.

28:04

И чтобы найти больше вещей, больше объектов, нужно включать, соответственно, ссылки на другие URI.

28:16

Англ.

28:21

Да, соответственно, и 5-звездочная модель данных.

28:25

Англ.

28:31

Соответственно, да, 5-звездочная модель реализации, создания сети, в которой наборы данных будут связаны между собой.

28:45

Англ.

28:47

И в этой модели две звезды присваиваются данным, которые позволяют повторное использование.

28:55

Англ.

28:58

Три звезды – данным, распространяемым по открытым стандартам.

29:03

Англ.

29:07

И четвертые и пятые звезды – это связь больших массивов данных.

29:15

Англ.

29:19

Давайте посмотрим на один пример связанного набора данных.

29:24

Англ.

29:30

Как я уже говорила, дипипедия – это версия википедии в виде связанных данных.

29:39

Англ.

29:45

Это коллективный проект по извлечению структурированной информации из википедии.

29:52

Англ.

29:56

И предоставление этой информации в доступном виде в сети.

30:00

Англ.

30:06

Соответственно, это позволяет пользователям задавать вопросы, касающиеся содержания википедии.

30:15

Англ.

30:19

И при этом ссылаться на другие источники сети.

30:24

Англ.

30:26

Например.

30:28

Англ.

30:35

Как мы видели уже, это URI Санкт-Петербурга.

30:39

Англ.

30:41

Или мы можем посмотреть на Москву.

30:48

Англ.

30:57

Эта страница соответствует принципам открытых, связанных данных.

31:04

Англ.

31:08

У нас есть уникальные идентификатор для этого города, для Москвы.

31:13

Англ.

31:19

При этом это URI http, мы можем посмотреть его в браузере.

31:23

Англ.

31:26

И пользователям предоставлена полезная информация.

31:31

Англ.

31:34

Здесь, например, какова площадь общая.

31:37

Англ.

31:40

В какой стране город находится.

31:42

Англ.

31:45

Имя его главы.

31:50

Англ.

31:53

И также ссылки на другие наборы данных.

31:57

Англ.

31:59

Например.

32:37

Англ.

32:43

Да, вот это ссылки, которые ведут на страницы самой дипипедии.

32:49

Англ.

32:56

Например, здесь мы видим, что в Москве родились следующие граждане.

33:01

Англ.

33:27

Проблема качества данных.

33:29

Англ.

33:31

Об этом мы поговорим в следующей части.

33:34

Англ.

33:38

Посмотрим на место жительства вот этого человека.

33:43

Англ.

33:48

Я могу просто кликнуть на эту ссылку и получу информацию об этом человеке.

33:53

Англ.

33:56

Подождем, пока закроется.

34:01

Англ.

34:08

Это полезная вещь, потому что у википедии в настоящий момент ограниченные возможности.

34:13

Англ.

34:17

Она не может ответить на простые вопросы, как например.

34:28

Англ.

34:33

Что общего между двумя городами, например, Инсбургом и Лейпцигом.

34:38

Англ.

34:45

Какие, кто меры центральных городов. Городов Центральной Европы, которые подняты более, чем на 1000 метров.

35:00

Англ.

35:04

В каких фильмах играют одновременно Бред Питт и Анджелина Джоли.

35:09

Англ.

35:22

И не может отвечать на комплексные вопросы, например, не может найти всех футболистов, которые играли, были вратарями в клубах, у которого есть стадион более, на более, чем 40 000 мест и, которые при этом родились в стране, в которой более 10 млн жителей.

35:52

Англ.

36:06

В этом смысле полезна дипипедия, потому что там можно сопоставить несколько параметров на основе данных.

36:18

Англ.

36:24

Мы подошли к завершению первой части презентации.

36:28

Англ.

36:34

Мы можем сделать небольшой перерыв на вопросы и ответы, а потом перейти ко второй части.

36:43

Ведущий:

Друзья, у кого есть вопросы?

36:59

Вопрос по-англ.

37:21

Тот же вопрос по-русски:

Мы хотим создать приложение, которое соответствует пяти звездам. Однако, встает вопрос: как мы может это сделать? Нам необходимо автоматически связать эти данные.

37:36

Ответ на вопрос по-англ.

37:40

Есть автоматические и полуавтоматические инструменты.

37:45

Англ.

37:54

Все эти инструменты могут конвертировать данные из форматов csv, xml в rdf.

38:01

Англ.

38:05

Во второй части моей презентации.

38:07

Англ.

38:10

Я покажу инструмент, который называется LODrefine.

38:21

Вопрос по-англ.

39:21

Тот же вопрос по-русски: соответственно, вопрос в том, что для создания связанных данных нам нужна четкая структура входных данных, даже если они в формате rdf для того, чтобы их правильно спарсить в rdf. Потому, как если мы спросим, то есть смотрим на обыкновенные данные, допустим даже не на статью википедии, где данные уже структурированы, то мы не можем определить Москва является местом жительства этого человека, местом рождения, или местом смерти, если у нас нет прямого линка на это.

39:54

Ответ на вопрос по-англ.

40:07

Если я понимаю Ваш вопрос правильно, Вы спрашиваете о том, как узнать, правильно ли сконвертированы данные.

40:16

Англ.

40:18

Это правильно?

40:20

Англ.

40:27

Это зависит от публикатора набора данных.

40:30

Англ.

40:32

И от самого источника данных.

40:34

Англ.

40:37

В случае с дипипедией.

40:39

Англ.

40:47

Если факт в википедии неверен.

40:51

Англ.

40:56

То тоже самое перейдет в дипипедию.

40:59

Англ.

41:03

Это как раз тот момент, когда начинается разговор о качестве данных.

41:09

Англ.

41:12

Когда факты нуждаются в подтверждении.

41:16

Англ.

41:25

Потому что связанные данные представляют это в структурированной форме.

41:29

Англ.

41:32

Вопрос по-англ.

42:00

Тот же вопрос по-русски: соответственно, как много атрибутов для распознования, что значат эти данные, требуется создать? То есть это один атрибут для места рождения, один атрибут для места смерти, один атрибут для места жительства. И встает вопрос: приходится ли нам создавать эти атрибуты самостоятельно?

42:22

Ответ по-англ.

42:28

Если Вы спрашиваете о количестве атрибутов, то их число в общем-то неограниченно.

42:35

Англ.

42:43

Настолько, насколько много фактов доступно на вашем исходном источнике.

42:48

Англ.

42:55

Столько же триплетов будет в Вашем наборе связанных данных.

43:02

Ведущий:

Подождите, давайте уже три вопроса. Чуть позже уже если время останется. Кто еще хочет задать?

43:15

Вопрос по-русски:

Добрый день. Вот вникая в жизнь и цикл связанных данных, хотел бы задеть тему антологии данных. Создание антологии, связь с существующими антологиями, как они у вас там происходят? Расскажите. Такой вопрос к примеру. Если данные брать: пУтин и путИна. Как он будет искать у вас?

43:36

Перевод вопроса ан англ.

43:54

Снова перевод тог же вопроса ан англ.

44:10

Ответ по-англ.

44:23

При конвертации набора данных сюда можно указать тип конкретного источника.

44:30

Англ.

44:33

Например, Москвы тип место.

44:39

Англ.

44:46

Точно также Вы можете указать, что Путин – это человек.

44:51

Англ.

44:57

И это, соответственно, будет определять, о чем ,собственно, URI.

45:04

Англ.

45:07

Повторение вопроса по-англ.

45:22

Ответ по-англ.

45:30

Насколько я понимаю, большинство антологий создаются людьми.

45:34

Англ.

45:41

Даже если они создаются автоматически, они должны утверждаться, проверяться людьми.

45:47

Англ.

45:52

Хотя бы, чтобы удостовериться в том, что Путин не окажется местом.

46:03

Ведущие:

Коллеги, у кого еще есть вопросы?

46:09

Вопрос по-русски:

Вопрос вот в чем. Есть ли какие-то возможности использовать данные внутри бинарных форматов? Например, как использовать, агрегировать те картинки, у которых есть, как уже говорили, в которых используется лицо Путина?

46:30

Перевод этого вопроса на англ.

47:18

Ответ на вопрос по-англ.

47:42

Да, это возможно, потому что картинке присваиваются определенные тэги.

47:52

Англ.

47:57

И дальше можно искать по тэгам.

48:01

Комментарий зрителя по-англ.

48:11

Ответ спикера.

48:13

Соответственно, без тэгов это использовать невозможно.

48:16

Англ.

48:19

Ведущие:

Коллеги, дальше вопросы. Вот.

48:23

Вопрос по-англ.

49:16

Первеод вопроса на русски:

Суть вопроса в том, для интернета, для людей, недоступность одного из сервисов не является проблемой. Ну, конечно, если это не гугл. В то же время, когда мы имеем связанную картину данных, которая сейчас изображена на экране, недоступность одного из центральных наборов данных, ставит под угрозу связанность всего массива данных. Вопрос: что с этим делать?

49:51

Ответ по-англ.

50:09

Если один из наборов данных недоступен в данный момент, то, конечно, вы не можете извлечь из него информацию, но как раз в этот момент.

50:24

Англ.

50:37

Но вы видите, что эти наборы данных связаны друг с другом.

50:42

Англ.

50:46

У них всегда несколько, много ссылок на разные наборы данных.

50:50

Англ.

51:00

Комментарий зрителя по-англ.

51:20

Соответственно, подводя итоги этой дискуссии открытая проблема состоит в том, что действительно, пока все в таком виде. Если какой-то набор данных недоступен, то в принципе можно получить информацию из другого набора данных. Но он может быть не идентичным, соответственно, этому, или не релевантным. Соответственно, следующая задача состоит в том, чтобы вероятно создавать зеркала таких же наборов там и так далее. Ну, как в сети, со страницами.

51:51

Ведущие:

Еще вопросы?

51:56

Вопрос по-русски:

Представьте себе программиста, которые решил использовать линковые данные и допустим он в своей программе обращается к упомянутой Вами странице Санкт-Петербурга. Обращается и хочет узнать, кто там родился. Но на сайте очень хорошо поработали и записали туда всех, родившихся в Санкт-Петербурге. Файл получился размером эдак мегабайт в 100. Система его, не справилась с таким объемом, программа зависла, и он решил больше не пользоваться линковыми данными. Такая ситуация возможна?

52:39

Перевод вопроса ан англ.

53:28

Уточнение вопроса спикером по-англ.

53:39

Ответ на вопрос по-агл.

53:45

Существует специфический язык запросов для связанных данных.

53:48

Англ.

53:50

Спаркл.

53:51

Англ.

53:59

И там вы можете указать в точности те элементы информации, которые Вам нужны.

54:06

Англ.

54:25

Вы просто хотите знать, какие люди родились в Санкт-Петербурге?

54:28

Допустим да, хотелось бы знать.

54:34

Англ.

54:41

Соответственно, можно указать лимит результатов, которые Вы хотите получить.

54:46

Англ.

54:55

И соответственно, так сказать, можно группами выдавать лимитированный ответ.

55:00

Ведущий:

Я немножко дополню к ответу, что действительно есть язык спаквиль и обращение к страницам. Оно сделано для удобства, в принципе, там все возможности пажинации, выбора, оно есть. Достаточно гибкий. Просто он там чуть более сложный. Просто обращение к страницам. Коллеги, еще вопросы? Тогда, переходим, наверно, ко второй части.

55:23

Англ.

55:25 – 56:15 – перерыв.

56:22

Англ.

56:30

Вторая часть моей презентации посвящена качеству связанных данных.

56:34

Англ.

56:40

Мы посмотрим, что означает качество данных именно в применении к связанным данным.

56:44

Англ.

56:46

Почему это важно.

56:47

Англ.

56:49

Какие сейчас есть проблемы.

56:51

Англ.

56:55

Направления и способы измерения качества данных.

56:58

Англ.

57:01

И мы посмотрим на инструменты для оценки в качестве примеров.

57:08

Англ.

57:10

Почему качество данных важно?

57:13

Англ.

57:19

Потому что во-первых, данные хороши настолько, насколько они качественны.

57:25

Англ.

57:32

Как мы видели недавно, как мы говорили, если Путину присвоено тип место.

57:37

Англ.

57:40

То пользователь не может использовать эти данные.

57:42

Англ.

57:48

То есть желательно знать, какие проблемы есть в наборе данных заранее.

57:52

Англ.

57:58

И чтобы потребитель мог полностью, в полной мере использовать доступную информацию.

58:08

Англ.

58:14

Соответственно, проблема. Первая – это измерение качества данных.

58:20

Англ.

58:23

Автоматически, или полуавтоматически.

58:27

Англ.

58:32

Определение способов измерения данных, качества данных. Качественные и количественные. Для каждого, соответственно, направления.

58:44

Англ.

58:47

Масштабируемость оценки.

58:50

Англ.

58:56

Показывать пользователю результаты этой оценки.

59:00

Англ.

59:04

И по итогам оценки исправлять качество данных.

59:09

Англ.

59:13

Давайте посмотрим на некоторые аспекты качества данных.

59:17

Англ.

59:24

Проблемы качества данных – это, собственно, комплекс проблем, которые могут оказывать воздействие на потенциал приложений, которые обрабатывают эти данные.

59:40

Англ.

59:52

Соответственно, методология оценивания качества данных. то есть оценивание того элемента данных, оценивание насколько определенный элемент, набор данных соответствует запросу пользователя в каждом конкретном случае.

01:00:18

Англ.

01:00:28

Соответственно, оценка качества включает оценивание по разным направлениям, и соответственно, способы измерения этого качества. И, соответственно, да, оценивает по критериям актуальности для запроса.

01:00:52

Англ.

01:00:58

И соответственно, способ измерения – это процедура измерения качества этих, по этим направлениям.

01:00:12

Англ.

01:01:26

Значит, да, соответственно, в недавнем исследовании было выявлено 23 направления.

01:01:36

Англ.

01:01:38

Да, и они были разделены на шесть групп.

01:01:40

Англ.

01:01:46

Например, доступность набора данных.

01:01:49

Англ.

01:01:54

Насколько этому набору данных можно доверять.

01:01:58

Англ.

01:02:01

Насколько часто он обновляется. Насколько он своевременен.

01:02:05

Англ.

01:02:11

Насколько верны его семантическая и синтаксическая представленность, представление.

01:02:19

Англ.

01:02:25

Уже существуют инструменты для оценки качества.

01:02:29

Англ.

01:02:33

Можно их разделить на три типа.

01:02:35

Англ.

01:02:38

Автоматические, полуавтоматические и ручные.

01:02:42

Англ.

01:02:50

Среди автоматических можно назвать элоудистат, который мы посмотрим позже.

01:02:55

Англ.

01:03:01

Он собирает основную статистическую информацию о наборе данных.

01:03:07

Англ.

01:03:12

Сколько у него классов. Сколько у него свойств.

01:03:14

Англ.

01:03:23

Линк-ю – еще один автоматический инструмент.

01:03:25

Англ.

01:03:33

Который рассматривает связи между наборами данных и оценивает точность.

01:03:42

Англ.

01:03:52

Они, эти инструменты, они автоматические, потому что пользователь не может определять, по каким направлениям там, то есть задавать принципы оценки.

01:04:05

Англ.

01:04:10

Полуавтоматические инструменты – это LODrefine.

01:04:14

Англ.

01:04:21

Я его уже упоминала, он также может использоваться для конвертации данных.

01:04:25

Англ.

01:04:30

Из форматов типа xml, csv в rdf.

01:04:34

Англ.

01:04:45

А также он может выдать, помогает оценить такие первичные параметры качества данных, как полноту.

01:04:54

Англ.

01:05:00

Еще есть инструмент, который называется флемингс тул, но он доступен только по-немецки.

01:05:07

Англ.

01:05:12

И есть, соответственно, ручные инструменты – сивее и одиклинсто.

01:05:20

Англ.

01:05:31

Они требуют большого участия человека, который определяет какие наборы данных будут оцениваться, по каким измерениям и так далее.

01:05:44

Англ.

01:05:48

Давайте посмотрим на LOD starts.

01:06:05

Англ.

01:06:12

LOD starts берет обычно надор данных в дейта хабе.

01:06:17

Англ.

01:06:21

И выдает базовую статистику этих наборов данных.

01:06:28

Англ.

01:06:32

Он также сообщает, в каких наборах данных есть проблемы.

01:06:37

Англ.

01:06:42

И приводит цифры, сколько там триплетов.

01:06:48

Англ.

01:06:50

Сколько типов.

01:06:51

Англ.

01:06:52

Классов.

01:06:53

Англ.

01:06:53

Свойств.

01:06:54

Англ.

01:06:56

Ссылок.

01:07:00

Англ.

01:07:08

Он также приводит некоторый обзор того, сколько словарей используется в этих наборах данных.

01:07:15

Англ.

01:07:19

Про классы, свойства, типы данных.

01:07:23

Англ.

01:07:33

Также он показывает, сколько языков представлено – английский, французский, немецкий.

01:07:41

Англ.

01:07:47

Также мы можем посмотреть на каждый набор данных.

01:07:49

Англ.

01:07:54

Красный цвет указывает, что там есть проблемы.

01:07:58

Англ.

01:08:03

Которые желтые, они пока обрабатываются.

01:08:07

Англ.

01:08:14

И также мы можем узнать.

01:08:38

Англ.

01:08:46

То есть это инструмент, который предоставляет базовую статистику по набору данных.

01:08:54

Англ.

01:08:58

Доступен по адресу: starts lod to you.

01:09:07

Англ.

01:09:11

Собственно, я хотела показать LODrefine.

01:09:27

Англ.

01:09:31

LODrefine – это версия проекта LOD на основе google refine.

01:10:41

Англ.

01:10:44

Вернемся, когда загрузится.

01:10:47

Англ.

01:10:54

Хочу показать одно короткое видео, как преобразовывать, как обрабатывать набор данных в LODrefine.

01:11:03

Англ.

01:11:08

И, собственно, пока оно грузится, я уже перехожу к завершению этой презентации, лекции.

01:11:15

Англ.

01:11:21

И подводя итоги, я хочу упомянуть некоторые достижения и проблемы в области связанных данных.

01:11:31

Англ.

01:11:38

Мы уже видели, что большое количество наборов данных было конвертировано в формат связанных данных.

01:11:47

Англ.

01:11:50

И эти данные содержат порядка 50 млрд фактов.

01:11:55

Англ.

01:11:59

Это энергичное, всемирное сообщество исследователей.

01:12:08

Англ.

01:12:13

И уже началась индустриализация этого процесса.

01:12:18

Англ.

01:12:25

То есть такие компании, как ББС, Нью-Йорк Тайм, Фейсбук, Гугл уже используют связанные данные.

01:12:32

Англ.

01:12:45

И также правительство открывает свои данные.

01:12:49

Англ.

01:12:53

Это уже сделали в США и Великобритании.

01:12:56

Англ.

01:13:02

Собственно, цель состоит в том, чтобы установить, чтобы создать связанные данные, как промежуточный этап на пути к семантик-вебу, то есть к семантической паутине.

01:13:19

Англ.

01:13:25

Однако при этих достижениях есть и некоторые проблемы.

01:13:28

Англ.

01:13:30

Во-первых, это качество данных.

01:13:32

Англ

01:13:36

То есть, например, проверка насколько верны связи между наборами данных.

01:13:40

Англ.

01:13:44

Насколько данные последовательны.

01:13:47

Англ.

01:13:50

Насколько они масштабируемы, насколько работают.

01:13:54

Англ.

01:13:57

И юзабилити, то есть насколько с ними удобно работать.

01:14:02

Англ.

01:14:05

Также вопрос потребления данных.

01:14:08

Англ.

01:14:14

То есть на зачаточных стадиях находятся такие вещи, как обработка больших массивов данных, скимамепинг – соответствие по схемам, отображение данных на схемы. Установление сообщений между данными ссылками.

01:14:51

Англ.

01:15:01

Ведущий:

Спасибо за выступление. Друзья, если есть вопросы, давайте, может быть коротенько, чтобы не выбиваться сильно из графика. Сделаем небольшой перерыв. Есть вопросы?

01:15:14

Вопрос по-английски.

01:15:19

Тот же вопрос по-русски:

Кто потребитель этих данных?

01:15:23

Ответ по-англ.

10:15:36

Ну, то есть в разных областях знания потребители у этих данных разные. Например, в науках о жизни, в бионауках там будет, соответственно врачи, например.

01:15:49

Англ.

01:15:54

В области, в госсфере это будут граждане, например.

01:16:01

Англ.

01:16:12

Это не только для профессионалов, это для всех.

01:16:15

Ведущий:

Можно я немножко да отвечу на вопрос. Я заходил на яндекс и там одна из команд использовала данные из двпедии для того, чтобы их интегрировать в проект фактов по городам. То есть в принципе, это может использовать любой, вопрос только обычно в качестве данных. Главное, чтобы они были актуальными, но в плане удобства это достаточно быстро и просто. Они это сделали буквально за час. Синтергировали и завизаровали, а дальше вопрос был в визуализации и придании этому формы. Еще вопросы?

01:16:47

Вопрос по-англ.

01:16:56

Перевод вопроса на русский:

Можете порекомендовать какие-нибудь спаркл-проекты, чтобы хранить там связанные данные?

01:17:04

Ответ по-англ.

01:17:16

Уточнение вопроса по-англ.

01:17:30

Ответ по-англ.

01:17:39

Ну да, есть. Соответственно, вопрос о платформах, где можно.

01:17:47

Англ.

01:17:55

Да, ответственно, вот, например, платформа.

01:18:02

Ведущий:

Я могу немножко да пояснить, что в принципе да есть он-то вики, есть семантик медиа вики, есть платформа виртуоза, там в зависимости от них. То есть решений довольно много, и насколько я знаю, большинство из них у коллег как раз на сайте они перечислены, в том числе на сайте двпедии.

01:18:18

Список википедии перечисление в таблице, есть вариант.

01:18:26

Ведущий:

Коллеги, ну, давайте тогда переходить к небольшому перерыву и дальше уже к следующему докладчику. Еще раз спасибо за выступление. Спасибо за выступление.