

Slide URL

<https://vu5.sfc.keio.ac.jp/slide/>

Web情報システム構成法

No.13 LOD

萩野 達也 (hagino@sfc.keio.ac.jp)

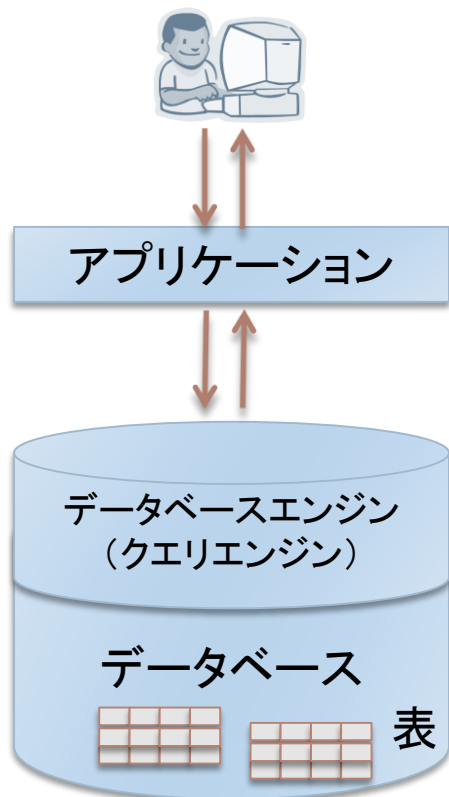
Semantic Webアプリケーションのアーキテクチャ

- ▶ システムの構成要素
 - ▶ RDFパーサ(解析)／シリアライザ(公開)
 - ▶ RDFデータストア
 - ▶ RDFのクエリエンジン
 - ▶ コンバータ／スクレイパ
 - ▶ 上記を利用したアプリケーション

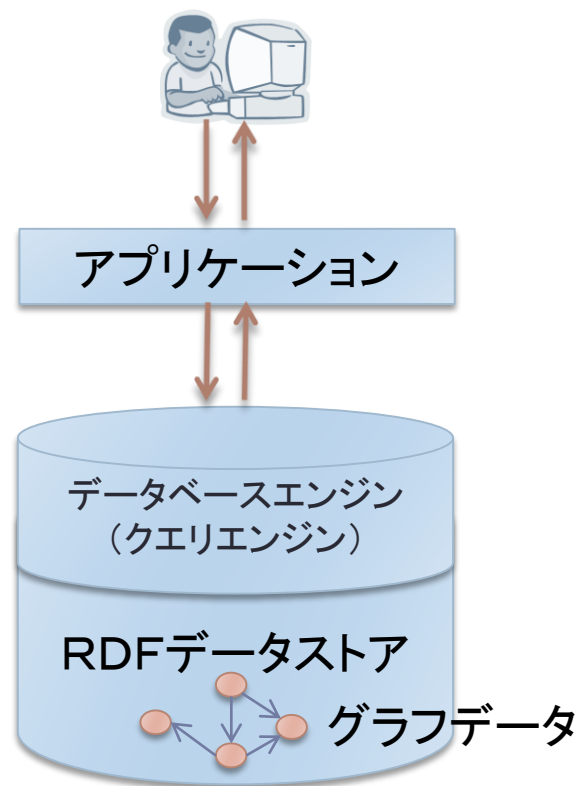
※ツールの多くは無償で入手できる

アーキテクチャ (1)

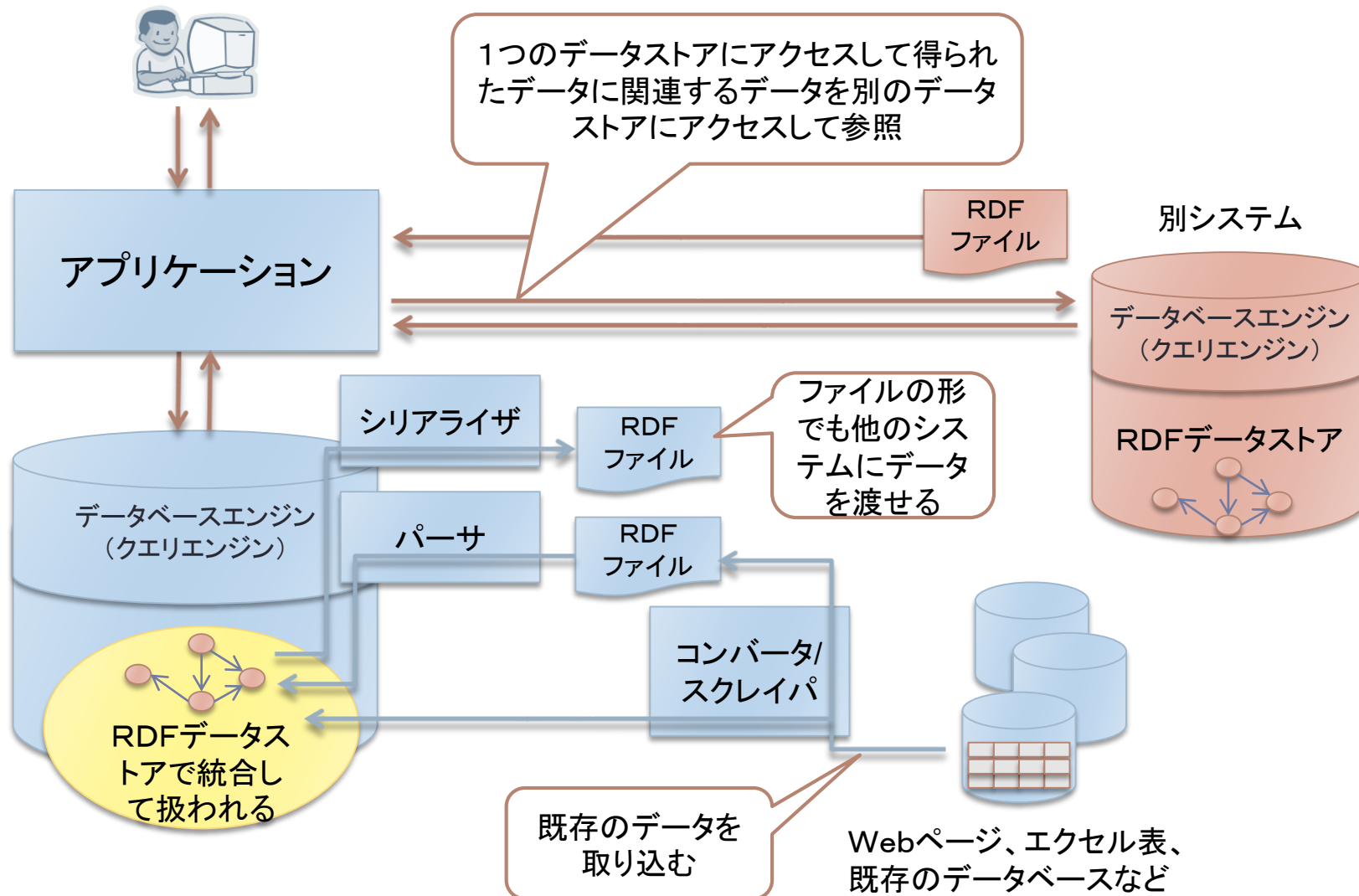
- ▶ データベースを用いた一般的なアプリケーションのアーキテクチャ



- ▶ RDFアプリケーションのアーキテクチャ



アーキテクチャ (2)



RDFパーサによるグラフ構造の解析

(a) fujisawa.rdf

```
<?xml version="1.0"?>  
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:s="http://example.jp/schema/">  
  <rdf:Description rdf:about="http://example.jp/doc.html">  
    <s:creator>  
      <rdf:Description>  
        <s:name>藤沢太郎</s:name>  
        <s:Email>taro@example...</s:Email>  
      </rdf:Description>  
    </s:creator>  
  </rdf:Description>  
</rdf:RDF>
```

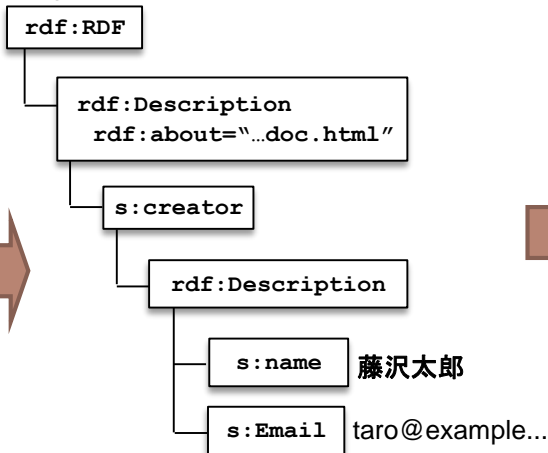
同じ意味を異なる表現で表したXMLで記述されたRDF

(b) fujisawa2.rdf

```
<?xml version="1.0"?>  
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:s="http://example.jp/schema/">  
  <rdf:Description rdf:about="http://example.jp/doc.html">  
    <s:creator s:name="藤沢太郎" s:Email="taro@example..." />  
  </rdf:Description>  
</rdf:RDF>
```

※シリアライズは逆にRDFグラフ構造のデータモデルからテキストを生成

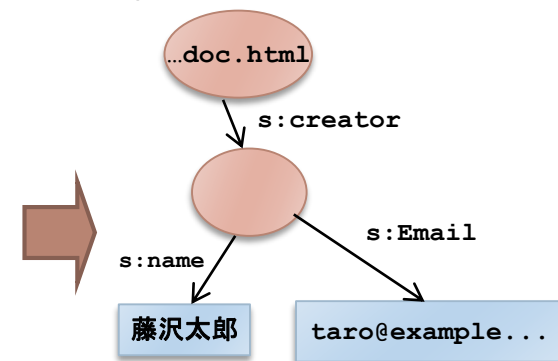
XMLパーサによる解析 (ツリー構造のデータモデル)



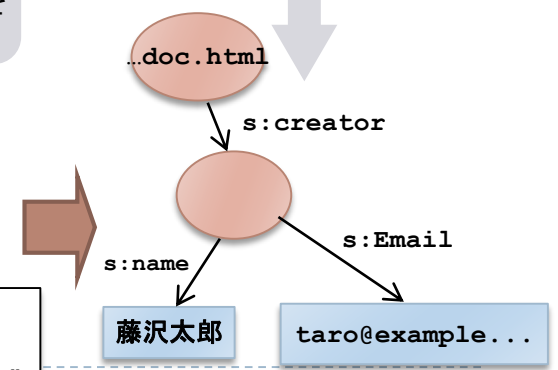
XMLパーサにかけたただけだと、要素の値や親子関係、属性の値を抽出することができるが、両者が同じ意味を持っているということを機械が判別できない



RDFパーサによる解析 (グラフ構造のデータモデル)

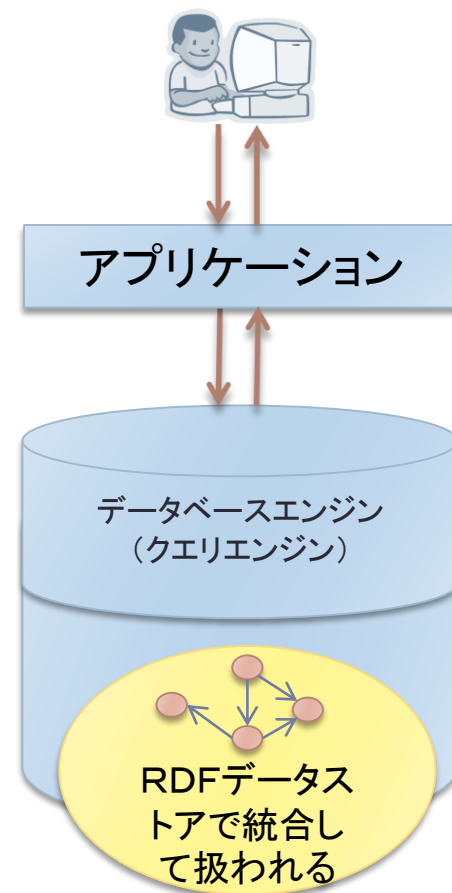


RDFパーサにかけると両者が同じ意味を持っていることを機械が判別できる



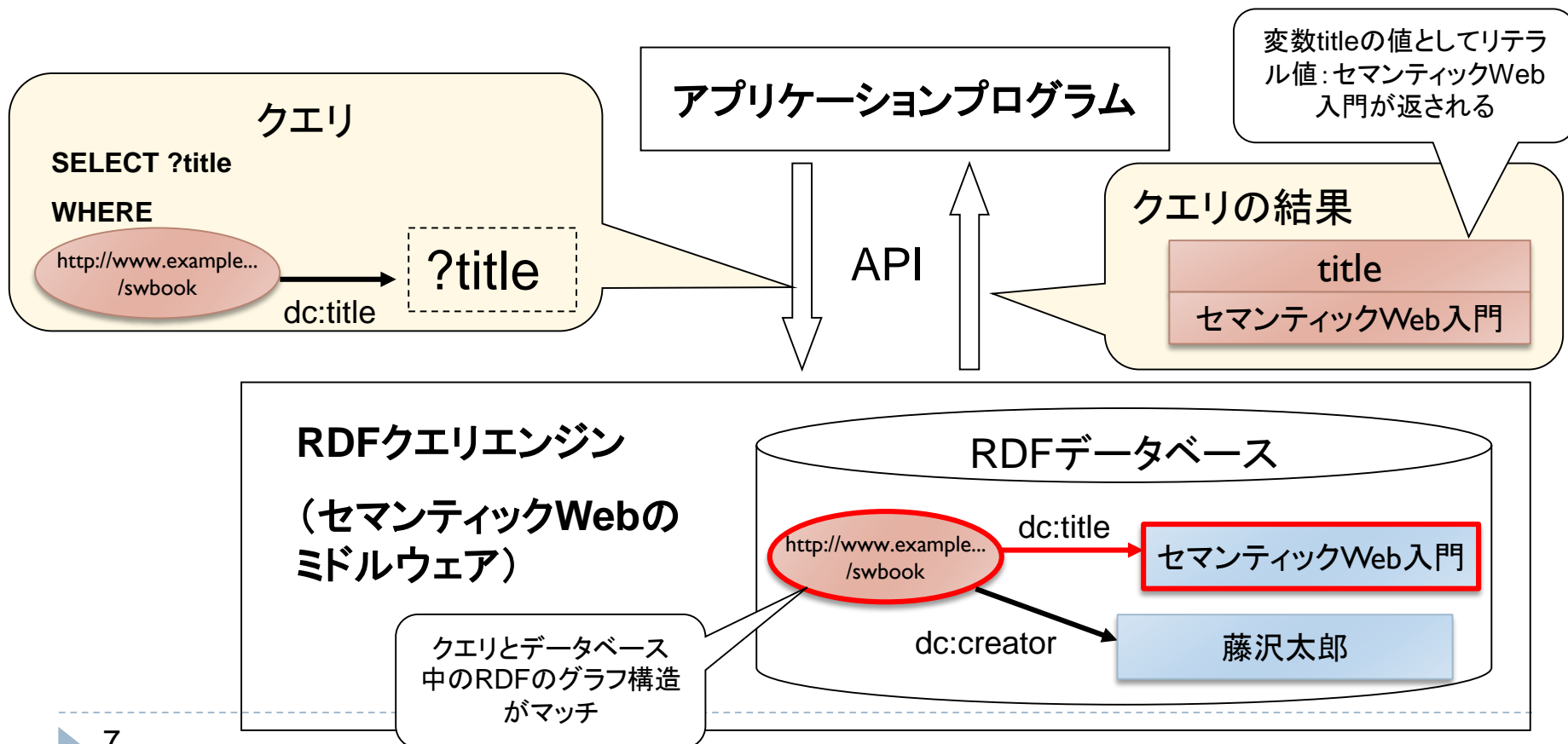
RDFのデータストア／データベース

- ▶ 格納されたRDFデータは統合される
- ▶ SPARQLなどのクエリ言語により検索することが可能になる
 - ▶ 複数のデータ集合の統合が基本機能として備わっている
 - ▶ リレーショナルデータベースの場合、通常、アプリケーション側でデータベースに格納されている形式に応じて問い合わせを行ない、結果をまとめる処理を行っている
 - ▶ RDFデータストアの場合、すべてのデータはRDF
- ▶ オープンソースから製品までさまざまなデータストアが公開されている



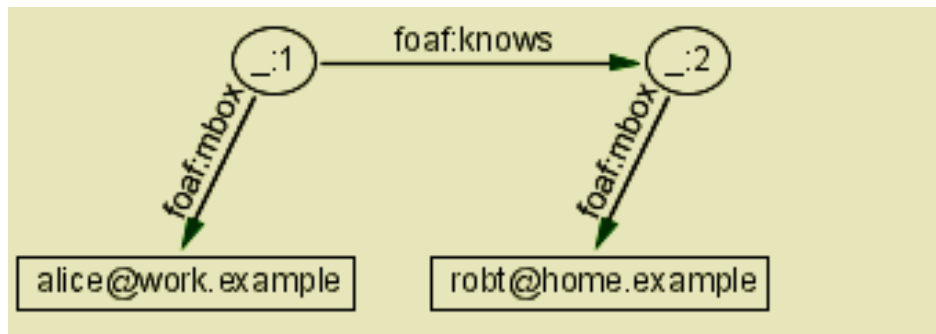
RDFのクエリ言語とは

- ▶ セマンティックWebのアプリケーションで利用されるRDFのクエリエンジンへのアクセス(入出力)を規定
 - ▶ データベースで管理されるグラフ構造のRDFデータからURIやリテラル値などの情報やサブグラフを取得



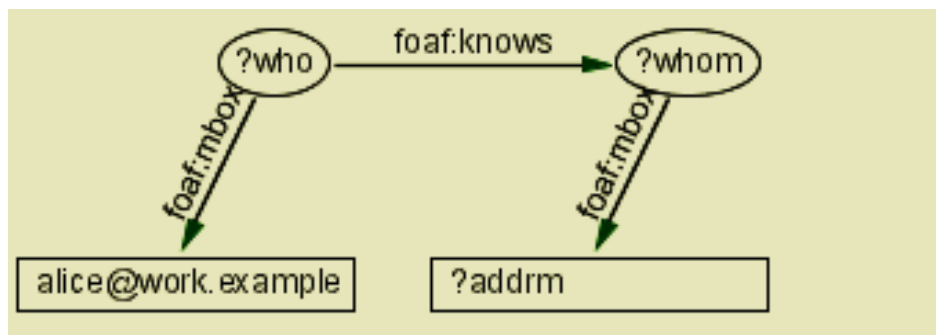
SPARQL

- ▶ SPARQL Protocol and RDF Query Language
- ▶ グラフベースのクエリ言語(トリプルパターンマッチング)



グラフ

(RDFデータベース内のデータ)



グラフパターン

(クエリ)

who	whom	addrm
_:1	_:2	"robt@home.example"

クエリ結果

(グラフパターンの変数 who, whom, addrmの値を提示)

SPARQLクエリ仕様 (複数マッチング)

- ▶ SELECT節
 - ▶ 値を取得したい変数の名前を記述
- ▶ WHERE節
 - ▶ 変数を含んだグラフパターン(トリプルパターン)

データ

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
_:a foaf:name "Johnny Lee Outlaw" .
_:a foaf:mbox <mailto:jlow@example.com> .
_:b foaf:name "Peter Goodguy" .
_:b foaf:mbox <mailto:peter@example.org> .
```

クエリ

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name, ?mbox
WHERE
{ ?x foaf:name ?name .
  ?x foaf:mbox ?mbox }
```

変数

トリプル
パターン

クエリ結果

name	mbox
"Johnny Lee Outlaw"	<mailto:jlow@example.com>
"Peter Goodguy"	<mailto:peter@example.org>

SPARQLクエリ仕様 (値の制約)

データ

```
@prefix dc: <http://purl.org/dc/terms/> .
@prefix : <http://example.org/book/> .
@prefix ns: <http://example.org/ns#> .

:book1 dc:title "SPARQL Tutorial" .
:book1 ns:price 42 .
:book2 dc:title "The Semantic Web" .
:book2 ns:price 23 .
```

クエリ

```
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX ns: <http://example.org/ns#>
SELECT ?title ?price
WHERE { ?x ns:price ?price .
  FILTER (?price < 30.5)
  ?x dc:title ?title . }
```

値の制約
(**price**の値
が30.5未満
のパターンのみ
マッチ)

クエリ結果

title	price
"The Semantic Web"	23

SPARQLクエリ仕様 (オプションルマッチング)

```
@prefix foaf:      <http://xmlns.com/foaf/0.1/> .
@prefix rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
_:a  rdf:type      foaf:Person .
_:a  foaf:name     "Alice" .
_:a  foaf:mbox     <mailto:alice@work.example> .

_:b  rdf:type      foaf:Person .
_:b  foaf:name     "Bob" .
```

データ

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE { ?x foaf:name ?name .
       OPTIONAL { ?x foaf:mbox ?mbox }
}
```

クエリ

クエリ結果

name	mbox
"Alice"	<mailto:alice@example.com>
"Bob"	

オプションルパターン(このパターンがマッチしなくても結果は得られる)

DBpediaのSPARQLエンドポイント

- ▶ DBpedia
 - ▶ Wikipediaから構造化データを抽出しWeb上で使えるようにしたもの
 - ▶ <http://dbpedia.org/>
- ▶ DBpediaデータにクエリを発行できるSPARQLエンドポイント
 - ▶ <http://dbpedia.org/sparql>
- ▶ 日本語版のDBpediaも公開された
 - ▶ <http://ja.dbpedia.org/>



実際にSPARQLクエリを発行して動作を確認してみたい

コンバータ / スクレイパ

▶ コンバータ

- ▶ 別の形式のデータをRDFに変換するツール
- ▶ 例: エクセルの表データ, 関係データベース

▶ スクレイパ

- ▶ 構造化された情報をWebページから抽出するツール

構造化データをWebページに埋め込む

- ▶ 構造化データ(メタデータ)をWebページに埋め込むための仕様
 - ▶ RDFa
 - ▶ <http://www.w3.org/TR/rdfa-core/>
 - ▶ Microdata
 - ▶ <http://www.w3.org/html/wg/drafts/microdata/master/>
 - ▶ microformats
 - ▶ <http://microformats.org/>

John knows

```
<a about="mailto:john@example.org"
  rel="foaf:knows"
  href="mailto:sue@example.org">Sue</a>.
```

HTMLに左記のようなRDFaの記述が埋め込まれている場合

Webブラウザで人間が見ると



John knows [Sue](#).

RDFaの仕様を解釈できる機械は以下のメタデータを読み取ることができる

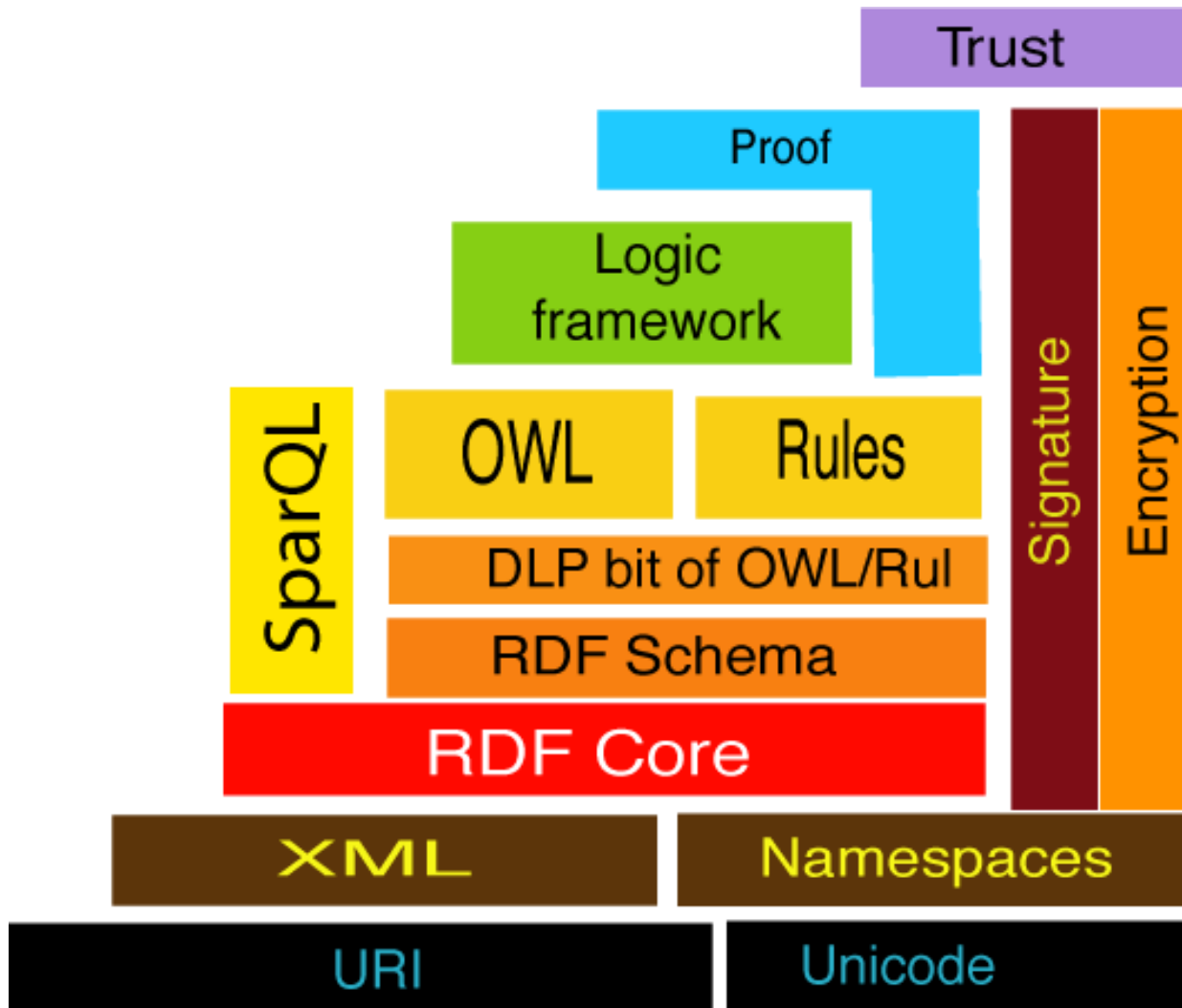


```
<mailto:john@example.org> foaf:knows <mailto:sue@example.org> .
```

Webページに埋め込まれた構造化データの活用

- ▶ 構造化データがWebページに記述されていると、ページを参照したアプリケーション(機械)がそれを利用できる
- ▶ Googleは収集したWebページの構造化データを読み取っている
 - ▶ 検索結果に表示する文字列(スニペット)といっしょに適切なプロパティ(属性)情報を表示させている
 - ▶ リッチ スニペット(microdata, microformats, RDFa, データ ハイライター)
 - ▶ <http://support.google.com/webmasters/bin/answer.py?hl=ja&answer=99170&topic=1088472&ctx=topic>
 - ▶ 構造化データテストツールを用いると、Webページに埋め込まれている構造化データを確認できる
 - ▶ <http://support.google.com/webmasters/bin/answer.py?hl=ja&answer=173839&topic=1088473&ctx=topic>

セマンティックWeb階層



Linked Data と Linked Open Data (LOD)

▶ オープンデータ

- ▶ 誰でもが利用して良いようにデータを公開する
- ▶ 政府のデータ, 製品情報, など

▶ Linked Data

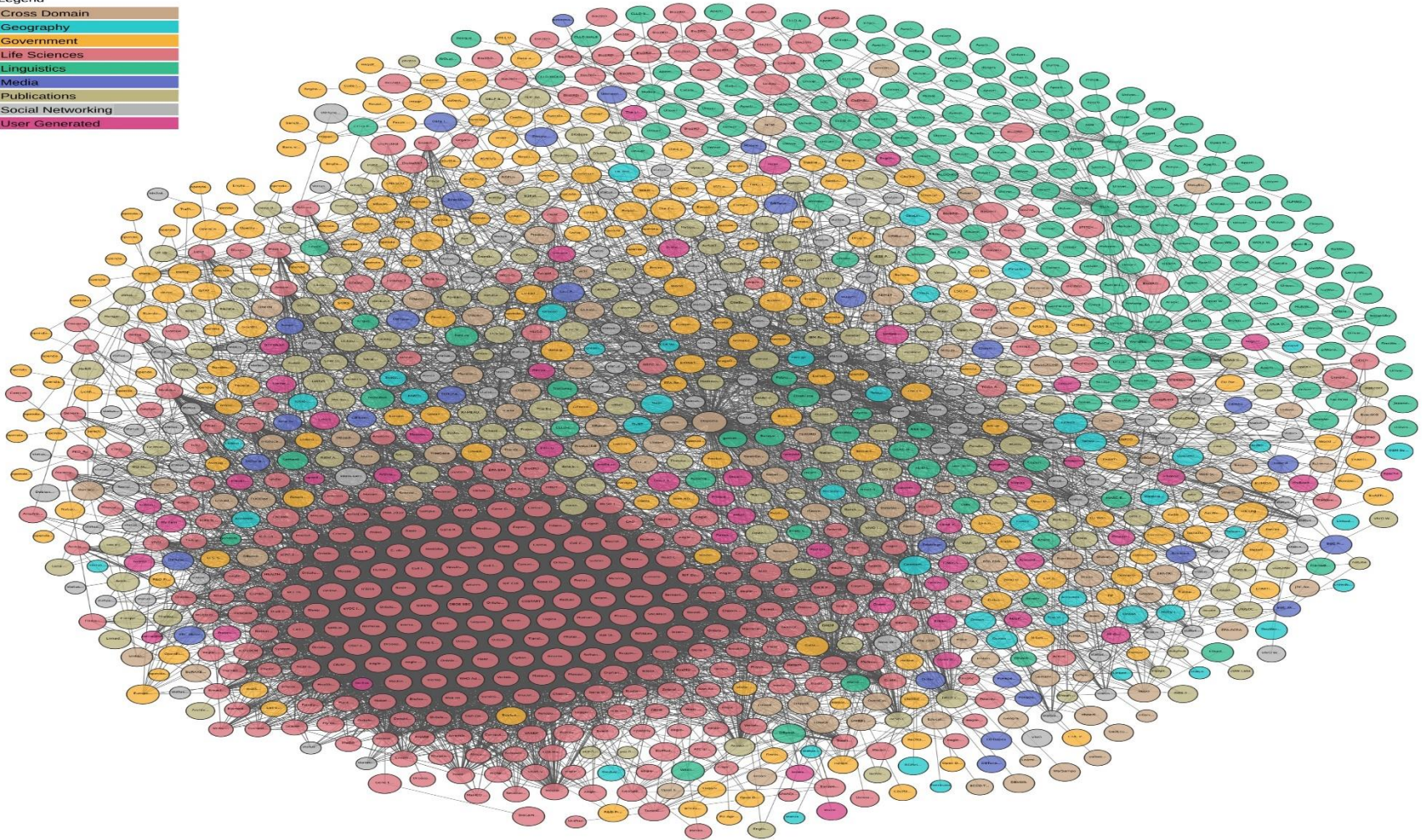
- ▶ 事物(資源)の識別子としてのURIを使う
- ▶ データ記述の標準フォーマットであるRDFを使う
- ▶ 共通のURIを使うことでデータ同士をリンクして活用する

▶ Linked Open Data

- ▶ オープンなライセンスの下で提供するLinked Data



Linked Open Data (LOD)



The Linked Open Data Cloud from lod-cloud.net



オープンデータ

- ▶ データが公開されているだけでなく、再利用を許可する
 - ▶ 公開されているので、だれでもが見ることができる
 - ▶ だれでも使うことができる
 - ▶ だれでもアプリケーションを作っても構わない

- ▶ 公共データ
 - ▶ 公共データは公開されるべきである
 - ▶ 政府が集めたデータ

オープンの定義 (version 1.1)

1. アクセス

作品の一部ではなく全てが、複製のための適正な価格あるいはインターネットによる無償ダウンロードにより提供されてなければなりません。また、作品は、変更可能で便利な形式で提供されなければいけません。

2. 再頒布

ライセンスは、作品自身あるいは様々な作品を集めたパッケージの一部として販売したり無償で頒布したりすることを制限してはいけません。ライセンスは、販売や頒布に関して使用料やその他の利用料を要求してはいけません。

3. 再利用

ライセンスは、作品の変更および派生した作品を作ることを許さなくてはいけません。また、それらの作品を元の作品と同じライセンスで頒布することを許さなければいけません。

4. 技術的制約の排除

作品は、上記に示した操作を行う場合に技術的な支障がない形式で提供されなければなりません。これは、仕様が公開され自由に利用可能で、料金や利用についての制限が課されてないオープンデータ形式を用いて作品を提供することによって達成することができます。

5. 帰属

ライセンスは、作品再頒布および再利用の条件として、作品の貢献者および作成者への帰属に関する要求してもかまいません。ただし、帰属に関する要求を行う場合には、煩雑でないようにしなくてはいけません。例えば、帰属情報を要求する場合、その帰属情報のリストは作品に付随しているべきです。

6. 完全性

ライセンスは、変更した作品を頒布する場合、元の作品と異なる名前やバージョン番号にすることを要求してもかまいません。

7. 個人やグループに対する差別の禁止

ライセンスは、特定の個人やグループを差別してはいけません。

8. 利用する分野に対する差別の禁止

ライセンスは、分野によって作品の利用を差別してはいけません。たとえば、企業での使用や遺伝子研究分野での使用についても制限をしてはいけません。

9. ライセンスの分配

作品に付随する権利は、その作品が再頒布された者全てに等しく認められなければならない、何らかの追加的ライセンスに同意することを必要としてはいけません。

10. 特定パッケージのみに制限するライセンスの禁止

作品に付与された権利は、それが特定のパッケージの一部であるということに依存するものであってはいけません。作品をパッケージから取り出したとしても、その作品のライセンスの範囲内で使用あるいは頒布される限り、作品が再頒布される全ての人々が、元のパッケージにおいて与えられていた権利と同等の権利を有することを保証しなければなりません。







11. 他の作品の頒布を制限するライセンスの禁止

ライセンスされた作品と共に頒布されている他の作品に制約を設けてはいけません。たとえば、同じ媒体で頒布されるすべての作品がオープンであることをライセンスは強制してはいけません。

クリエイティブ・コモンズ

▶ Creative Commons

- ▶ インターネット上の創作物に関する著作権
- ▶ 主に、文書、音楽、画像、動画などが対象

<p>CC BY</p> <p>表示</p> 	<p>原作者のクレジット(氏名、作品タイトルなど)を表示することを主な条件とし、改変はもちろん、営利目的での二次利用も許可される最も自由度の高いCCライセンス。</p>	<p>CC BY-SA</p> <p>表示—継承</p> 	<p>原作者のクレジット(氏名、作品タイトルなど)を表示し、改変した場合には元の作品と同じCCライセンス(このライセンス)で公開することを主な条件に、営利目的での二次利用も許可されるCCライセンス。</p>
<p>CC BY-ND</p> <p>表示—改変禁止</p> 	<p>原作者のクレジット(氏名、作品タイトルなど)を表示し、かつ元の作品を改変しないことを主な条件に、営利目的での利用(転載、コピー、共有)が行えるCCライセンス。</p>	<p>CC BY-NC</p> <p>表示—非営利</p> 	<p>原作者のクレジット(氏名、作品タイトルなど)を表示し、かつ非営利目的であることを主な条件に、改変したり再配布したりすることができるCCライセンス。</p>
<p>CC BY-NC-SA</p> <p>表示—非営利—継承</p> 	<p>原作者のクレジット(氏名、作品タイトルなど)を表示し、かつ非営利目的に限り、また改変を行った際には元の作品と同じ組み合わせのCCライセンスで公開することを主な条件に、改変したり再配布したりすることができるCCライセンス。</p>	<p>CC BY-NC-ND</p> <p>表示—非営利—改変禁止</p> 	<p>原作者のクレジット(氏名、作品タイトルなど)を表示し、かつ非営利目的であり、そして元の作品を改変しないことを主な条件に、作品を自由に再配布できるCCライセンス。</p>

クリエイティブ・コモンズ vs オープンの定義

▶ オープンの定義

- ▶ データに対する定義
 - ▶ 通常の著作権はデータには及ばない
 - ▶ 公開・再配布可能なデータのみが対象
 - ▶ フリーソフトのライセンスに近い
 - ▶ オープンソース
 - 一般的な無料ソフトの名称
 - ▶ フリーソフト
 - フリー(無料)であることを継承させる
 - 商用利用でお金を取ることができない
 - ▶ パブリックドメイン
 - 著作権や商標・特許などを放棄
 - 商用利用も可能になる

政府のオープンデータ

▶ アメリカ

▶ Open Government Initiative

- ▶ オバマ大統領「これまでにないオープンな政府を作る」
- ▶ 政府は透明であり, 参加可能であり, 共同しなくてはならない.

▶ Data.gov

- ▶ <https://www.data.gov/>
- ▶ 141,192データセット(2015年6月28日現在)

▶ ヨーロッパ

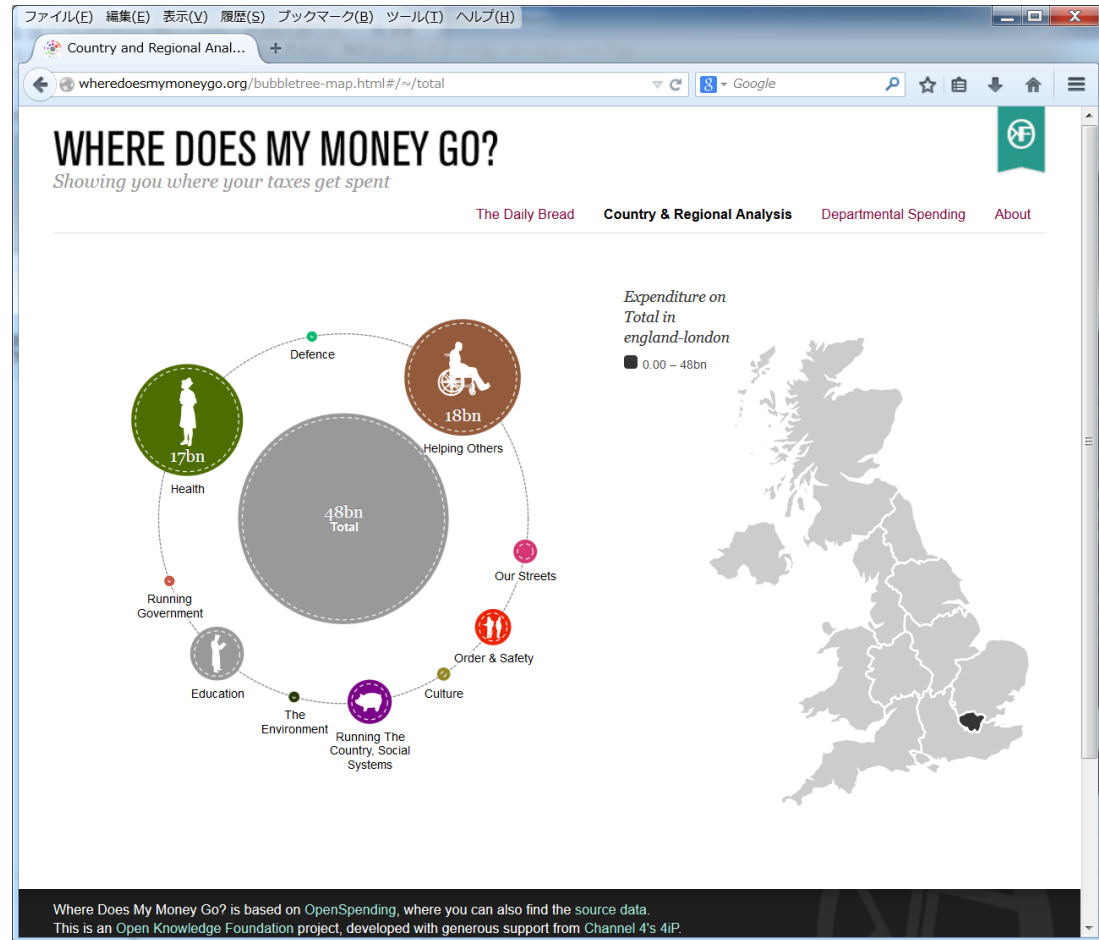
▶ [Data.gov.uk](https://data.gov.uk/)

▶ 日本

▶ www.data.go.jp

税金はどこへ行った？

- ▶ 政府がどのようなところにお金を使っているかを知るプロジェクト
- ▶ 政府の公開データを利用して市民が作成
- ▶ UK
 - ▶ <http://wheredoesmymoneygo.org/>
- ▶ Japan
 - ▶ <http://spending.jp/>



ゴミ収集に関するデータ

- ▶ いつ、どのようなごみ収集が行われているのか？
- ▶ <http://5374.jp/>
- ▶ Code for Kanazawa

The screenshot shows a web browser displaying the 5374.jp website. The page title is "5374.jp" and the URL is "5374.jp/en/". The website content includes the heading "5374.jp" and the text "“When will which garbage be collected?”". Below this, it states: "Garbage collection is becoming a serious problem. Code for Kanazawa is firstly focused on the correct way to dispose of garbage. With this app, if you move to a new area of Kanazawa, you can find information about different trash disposal times for your neighborhood." To the right, there are two smartphone screens displaying the app interface. The left screen shows a list of trash categories with their collection dates: "今日" (Today) for "資源" (Resources), "明日" (Tomorrow) for "燃やす" (Burnable), "7日後" (7 days later) for "燃やさない" (Non-burnable), and "23日後" (23 days later) for "びん" (Bottle). The right screen shows a detailed view of the "びん" category. Below the smartphone screens, there is a section titled "“Using 5374”" with the text: "Each type of trash is displayed in a different color. For trash that can be thrown away, tap the trash category to see the specific items that can be discarded. Set your location to automatically set your collection dates for each category of disposable trash. In the future, we are planning to use Smart Phone's GPS to find your home's position." At the bottom of the page, there is a purple bar with the text "“Update”" and "Information about 5374.jp" and a right-pointing arrow.

オープンデータの5つ星

▶ <http://5stardata.info/>

★ オープンライセンスでWebでデータ(形式はなんでもよい)を公開する

★★ 構造化データ(スキャナの画像ではなくExcelなどを使う)を公開する

★★★ 公開された形式でデータを公開する(ExcelではなくCSVにする)

★★★★ URIを使って表現し, 他のデータから参照可能にする

★★★★★ 他のデータにリンクする



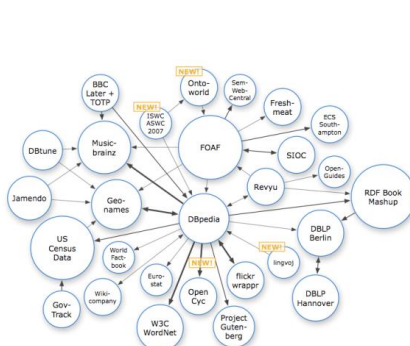
Linked Open Data

▶ Linked Data (<http://www.w3.org/DesignIssues/LinkedData.html>)

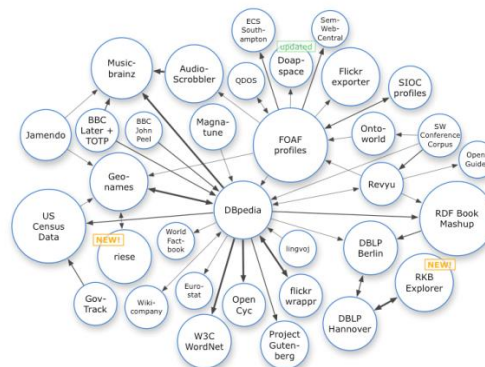
1. URIを使ってモノを表す
2. HTTP URIを使う (HTTPプロトコルでアクセス可能にする)
3. URIについて調べたときに、そのモノに関する情報をRDFあるいはSPARQLエンドポイントとして提供する
4. 他のURIに関する情報を含めることで、そのモノに関するさらなる情報の発見を可能にする



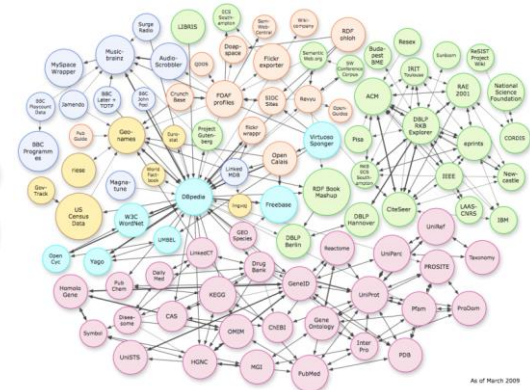
2007/5/1



2007/11/7



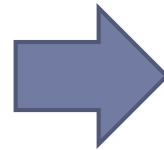
2008/3/31



2009/3/27

データをつなぐことの効果

2008年に弁護士がオハイオ州のゼーンズビルの水道の整備地図を作成



白人の居住状況とマッシュアップ

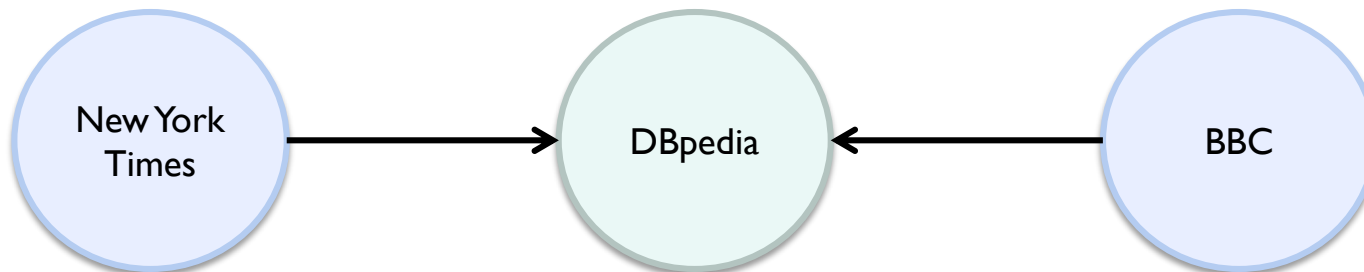
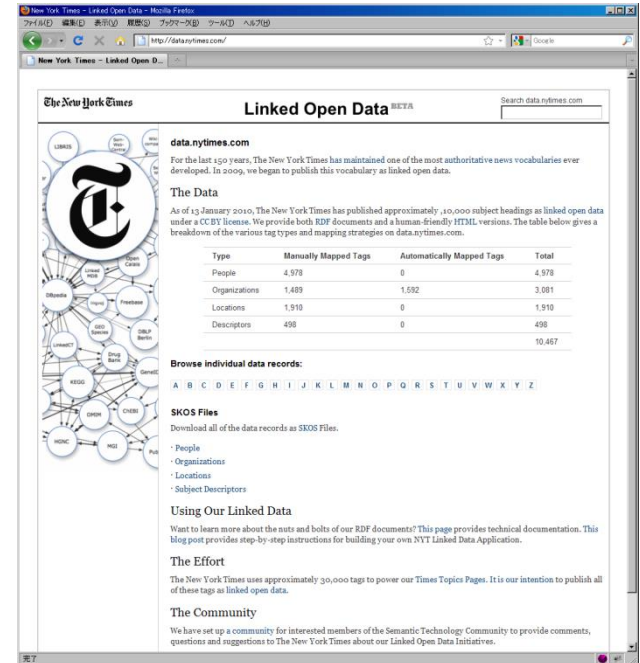


黒人多い居住地に差別的に水道が整備されていないことが分かり、市が10.9億円の賠償を命じられる

http://www.dispatch.com/live/content/local_news/stories/2008/07/11/Coal_Run.ART_ART_07-11-08_A1_SPANNJH.html

メディアデータの連携

- ▶ LODによってメディアデータを統合することが可能
 - ▶ ニューヨーク・タイムズ
 - ▶ <http://data.nytimes.com/>
 - ▶ 新聞記事(人・組織・場所)
 - ▶ BBC
 - ▶ 番組情報(BBC/Program)
 - ▶ 音楽情報(BBC/Music)
 - ▶ DBpedia
 - ▶ データのハブ



課題:LODの活用を考えよう

- ▶ 自分の身近にあるデータを組み合わせることで、どんな新しいアプリケーションを作ることができるかを提案しなさい。
 - ▶ 身の回りにあるデータを探しましょう。
 - ▶ 大学, 自宅, バイト先などにあるデータを探しましょう。
- ▶ 提出
 - ▶ <https://vu5.sfc.keio.ac.jp/kadai/>
 - ▶ 提案内容をテキストとして書きなさい
 - ▶ 締め切り: 7月21日正午

まとめ

▶ Linked Open Data

- ▶ RDFのデータをオープンなライセンスの下に公開し、異なるデータ同士をリンクして活用する試みとして注目を集める

▶ セマンティックWebアーキテクチャ

- ▶ RDFパーサ
- ▶ RDFデータストア
- ▶ SPARQL検索