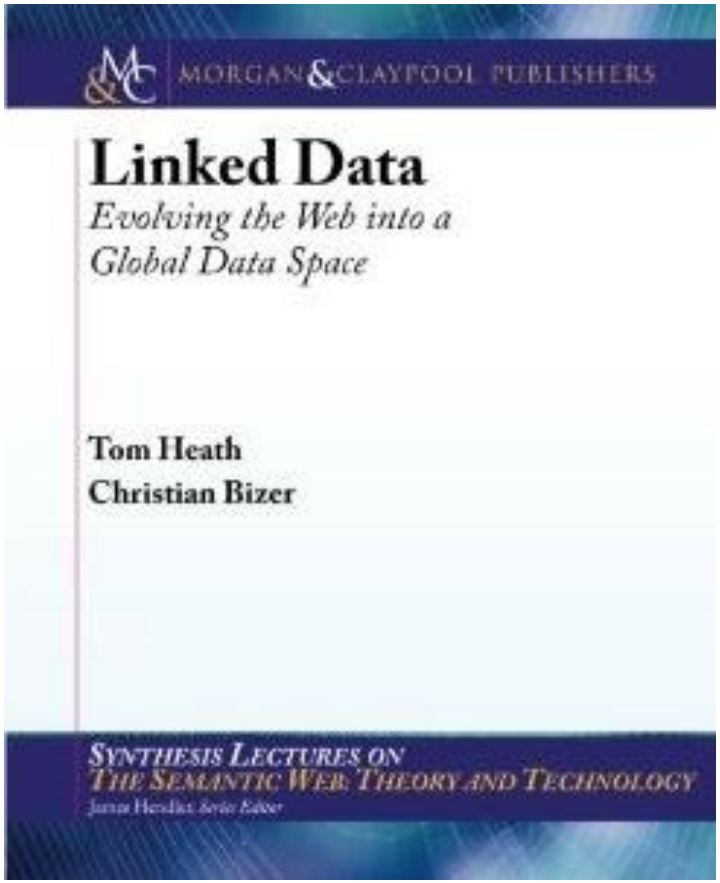
Linked Data

Erdoğan Doğdu

2012

Book

- Tom Heath and Christian Bizer (2011) ***Linked Data: Evolving the Web into a Global Data Space*** (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.
- http://linkeddatabook.com/book



Content

- **Principles** of Linked Data
- The Web of Data
- Linked Data **Design Considerations**
- Recipes for **Publishing** Linked Data
- **Consuming** Linked Data

Scenario

- BigLynx Productions
  - (fictional) independent television production company specializing in wildlife documentaries
- http://biglynx.co.uk/
  - information about the **company's goals and structure**
  - **profiles** of the permanent staff and of freelancers
  - listings of **vacancies for freelancers** to work on specific contracts
  - listings of **productions** that have been broadcast by the commissioning network
  - a **blog** where staff post news items of interest to the television networks and/or freelancers

Principles

- a set of **best practices for publishing and interlinking structured data** on the Web
- *Linked Data principles*
  - Tim Berners-Lee. Linked Data - Design Issues, **2006**. http://www.w3.org/DesignIssues/LinkedData.html

Principles

1. **Use URIs as names** for things.
2. **Use HTTP URIs**, so that people can look up those names.
3. When someone looks up a URI, **provide useful information**, using the standards (RDF, SPARQL).
4. Include **links to other URIs**, so that they can discover more things.

Idea

- Apply the general architecture of WWW to **sharing structured data on global scale**.
- WWW architecture
  - **URI**: globally unique identification mechanism
  - **HTTP**: universal access mechanism
  - **HTML**: widely used content format
  - **Hyperlinks** between Web documents on different servers

Idea

- WWW: *single global information space*
- *Web of Data: single global data space*
  (sharing data on global scale)

Principles

- (1) URI
  - WWW: Identify web documents and digital content
  - WofD: Identify real world objects and abstract concepts, e.g. people, places, cars, etc.
- (2) HTTP URIs
  - WWW: Mechanism to retrieve web resources
  - WofD: Mechanism to retrieve the description of objects/abstract concepts (dereferenced URIs over HTTP protocol)

Principles (ctd.)

- (3) Universal (data) representation
  - WWW: HTML serves the general representation format
  - WofD: RDF is the standard data model
- (4) Hyperlinks
  - WWW: Connecting web documents
  - WofD: Connecting data
- In summary
  - Extend the web to a global data space

Use of WoD

- Some examples
  - **Generic applications** to operate on data space
    - E.g. Linked data browsers
  - **Linked data search engines**
    - Crawling web of data to answer sophisticated queries
  - …

Naming things w/ URIs

- Way of **publishing data** on the web
- Use **HTTP URIs** because
  1. Globally unique names
  2. Means for accessing info describing the named entity

Make URIs dereferenceable

- HTTP URI should be **dereferenceable**
  - HTTP clients can look up the URI using the HTTP protocol and retrieve a description of the resource that is identified by the URI
  - **Descriptions** that are intended to be read by **humans** are often represented as **HTML**. Descriptions that are intended for consumption by **machines** are represented as **RDF** data.

Example


http://biglynx.co.uk/people/matt-briggs
http://biglynx.co.uk/people/scott-miller
http://xmlns.com/foaf/0.1/knows
http://biglynx.co.uk/people/linda-meyer

URIs to identify objects vs. docs describing them

- Objects vs. documents describing those objects
- Use different URIs
  - URI to identify object
  - URI to identify the document describing that object
- HTTP **content negotiation**
  - HTTP client sending in HTTP header what kind of document it prefers (HTML vs. RDF)

303 URIs vs. hash URIs

- Two strategies to name objects vs. docs describing them
  - 303 URIs
  - Hash URIs
- Leo Sauermann and Richard Cyganiak. Cool uris for the semantic web - w3c interest group note. http://www.w3.org/TR/cooluris/, 2008

303 URIs strategy

- Instead of sending the object itself over the network, the server responds to the client with the HTTP response code ***303 See Other*** and the URI of a Web document which describes the real-world object.
  This is called a ***303 redirect***.

303 See Other

- Example: **Person Dave Smith** (real-world object)
- Need 3 URIs:
  - http://biglynx.co.uk/people/dave-smith
    (URI identifying the **person Dave Smith**)
  - http://biglynx.co.uk/people/dave-smith.rdf
    (URI identifying the **RDF/XML document** describing Dave Smith)
  - http://biglynx.co.uk/people/dave-smith.html
    (URI identifying the **HTML document** describing Dave Smith)

Step 1

Request:

GET /people/dave-smith HTTP/1.1
Host: biglynx.co.uk
Accept: text/html;q=0.5, **application/rdf+xml**

Response:

HTTP/1.1 **303 See Other**
Location: http://biglynx.co.uk/people/dave-smith.rdf
Vary: Accept

Step 2

Request:

GET /people/**dave-smith.rdf** HTTP/1.1
Host: biglynx.co.uk
Accept: text/html;q=0.5, **application/rdf+xml**

Response:

HTTP/1.1 200 OK
Content-Type: **application/rdf+xml**

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:foaf="http://xmlns.com/foaf/0.1/">

<rdf:Description rdf:about="http://biglynx.co.uk/people/dave-smith">
<rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
<foaf:name>Dave Smith</foaf:name>
...
```

Hash URIs

- # fragment identifier
- To describe more items in the same document
- Big Lynx example: Two terms
  - http://biglynx.co.uk/vocab/sme#**SmallMediumEnterprise**
  - http://biglynx.co.uk/vocab/sme#**Team**
  - To retrieve:
    - GET **/vocab/sme** HTTP/1.1
      Host: biglynx.co.uk
      Accept: application/rdf+xml

Hash URI example

- **Request**

  GET **/vocab/sme** HTTP/1.1
  Host: biglynx.co.uk
  Accept: application/rdf+xml

- **Response**

  HTTP/1.1 200 OK
  Content-Type: application/rdf+xml;charset=utf-8

```
 <?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<rdf:Description rdf:about="http://biglynx.co.uk/vocab/sme#SmallMediumEnterprise">
        <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class" />
</rdf:Description>
<rdf:Description rdf:about="http://biglynx.co.uk/vocab/sme#Team">
        <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class" />
</rdf:Description>
...
```

Hash vs. 303

- Hash
  - Reduced number of HTTP calls
  - Returns all fragments (#parts)
  - Also used in RDFa (rdfa about attr)
- 303
  - More flexible; return seperate doc for each resource or a large doc for all
- 303 used more often to describe resources (large number of them such as Dbpedia concepts, 3.6M)
- Hash used more often to describe RDF vocabs (small set)
- Or a combination of 303 and hash is a better choice.

“Linked Data” data format

- RDF, Resource Description Format
  - Frank Manola and Eric Miller. RDF Primer. W3C, http://www.w3c.org/TR/rdf-primer/, February 2004
- Standard content format
- Triples: subject + predicate + object
  - Ex: “Matt Briggs” + “has nick name” + “Matty”

Triples

- Two types of triples:
  - **Literal triples**: URI + URI + **literal**
    - Describe properties of resources. E.g. a person's nick name
    - http://biglynx.co.uk/people/matt-briggs http://xmlns.com/foaf/0.1/nick "**Matty**"
  - **RDF links**: URI + URI + URI
    - Linking resources. E.g. A person knows another person
    - http://biglynx.co.uk/people/matt-briggs http://xmlns.com/foaf/0.1/knows http://biglynx.co.uk/people/dave-smith

RDF Serialization Formats

- RDF/XML
- RDFa
- Turtle
- N-Triples
- RDF/JSON

RDF/XML

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <rdf:RDF
3 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4 xmlns:foaf="http://xmlns.com/foaf/0.1/">
5
6 <rdf:Description rdf:about="http://biglynx.co.uk/people/dave-smith">
7 <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
8 <foaf:name>Dave Smith</foaf:name>
9 </rdf:Description>
10
11 </rdf:RDF>
```

RDFa

- 1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
  "http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
  2 <html xmlns="http://www.w3.org/1999/xhtml"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/">
  3
  4 <head>
  5 <meta http-equiv="Content-Type" content="application/xhtml+xml;
  charset=UTF-8"/>
  6 <title>Profile Page for Dave Smith
  7 </head>
  8
  9 <body>
  10
  11 Dave Smith
  12
  13 </body>
  14
  15 </html>

Turtle

```
1 @prefix rdf: <http://www.w3.org/1999/02/22-
    rdf-syntax-ns#> .
2 @prefix foaf: <http://xmlns.com/foaf/0.1/> .
3
4 http://biglynx.co.uk/people/dave-smith
5 rdf:type foaf:Person ;
6 foaf:name "Dave Smith" .
```

N-Triples

1 <http://biglynx.co.uk/people/dave-smith> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .

2 <http://biglynx.co.uk/people/dave-smith> <http://xmlns.com/foaf/0.1/name> "Dave Smith" .

RDF/JSON

- Not standard yet
- Keith Alexander. Rdf in json. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web*, 2008.

Web of Data

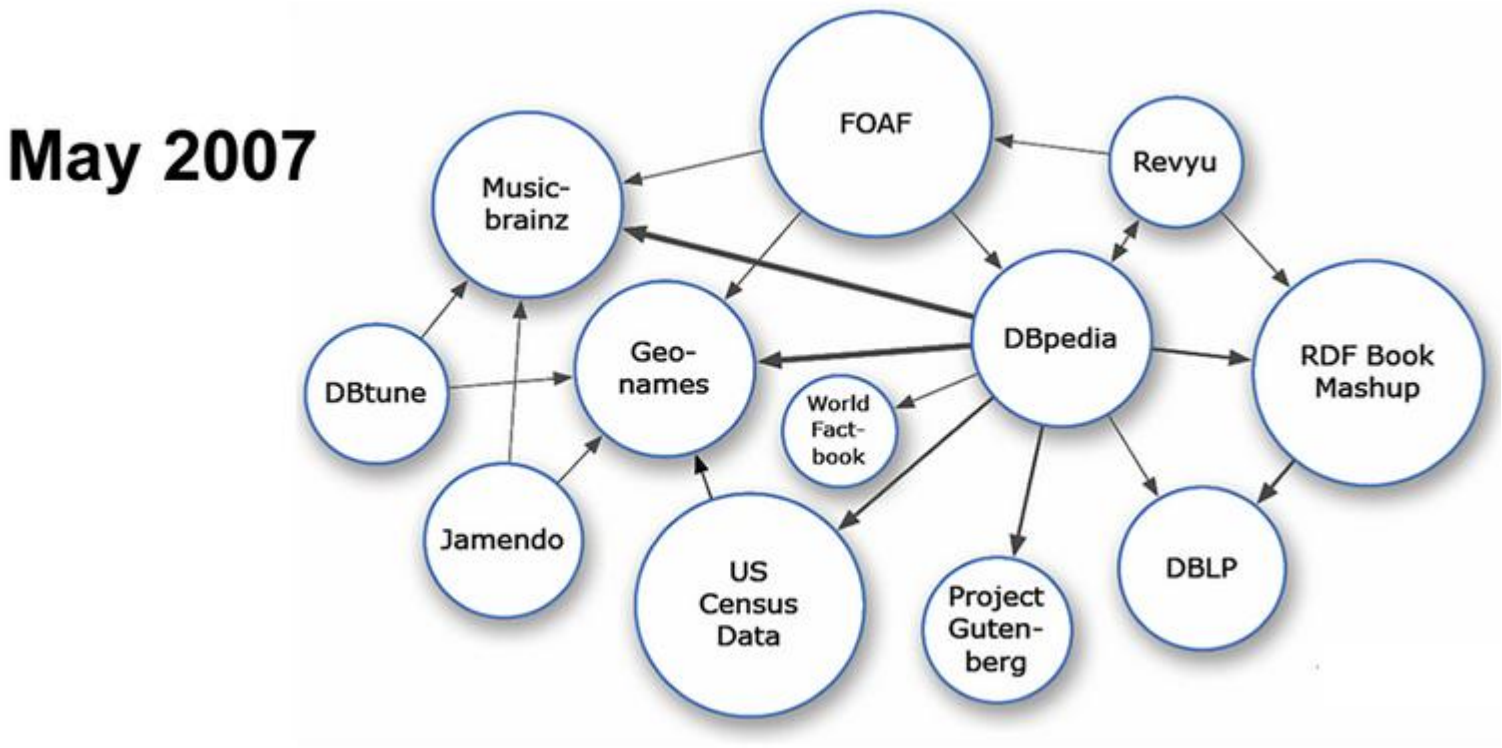
- Giant global data graph
  - consisting of **billions of RDF statements** from numerous sources
  - covering **all sorts of topics**, such as geographic locations, people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, statistical data, census results, online communities and reviews, etc.

Web of Data

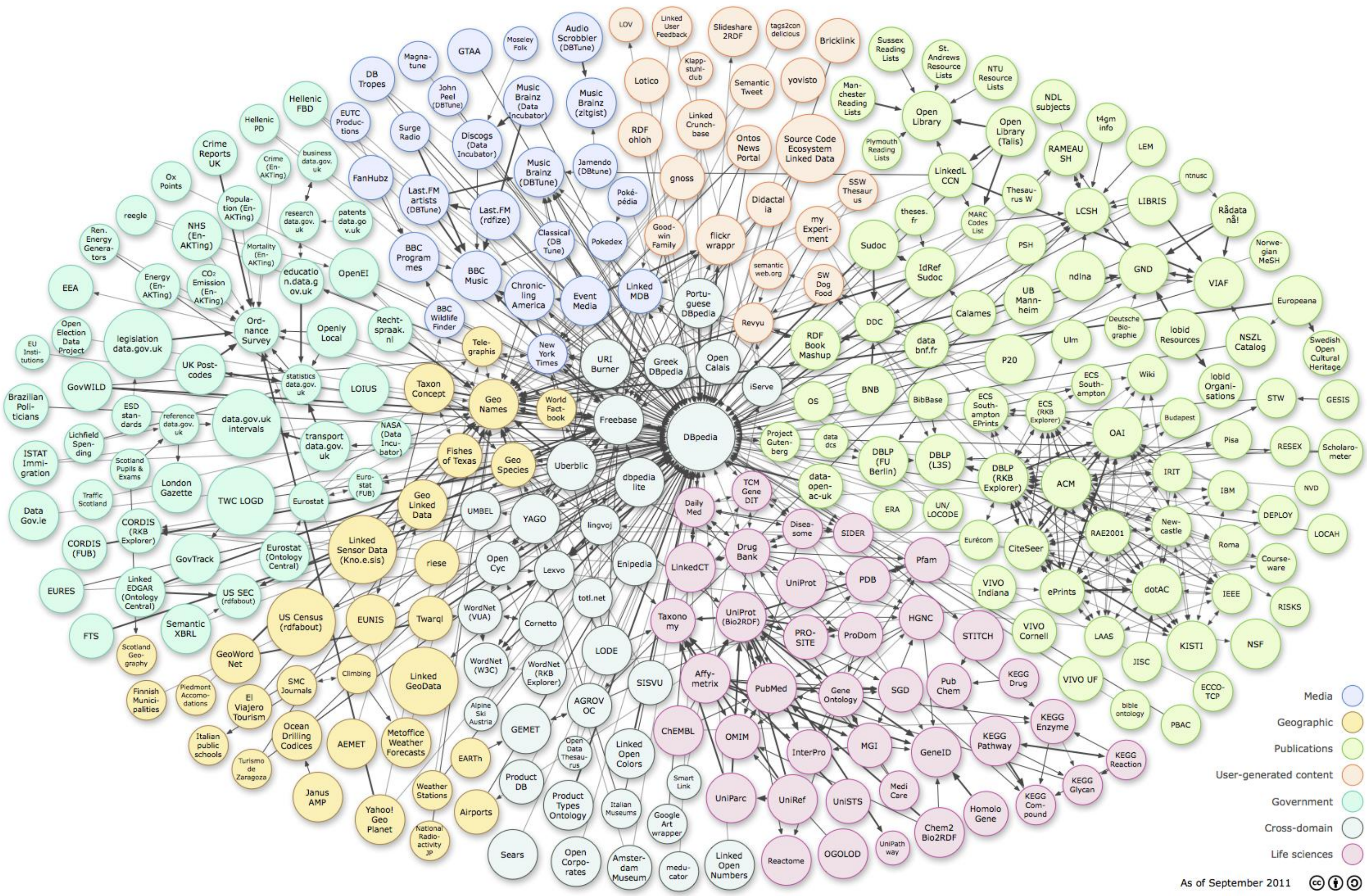
- Generic, can contain any type of data
- Anyone can publish
- Can contain contradictory information
- No fixed data sets, follow the links …
- No constraints on the choice of vocabs
- Data is self-describing (dereferenced)
- Standardized data access mechanism HTTP, standardized data model RDF simplifies data access (compare to web APIs).

On the web …

- W3C *Linking Open Data (LOD) project*



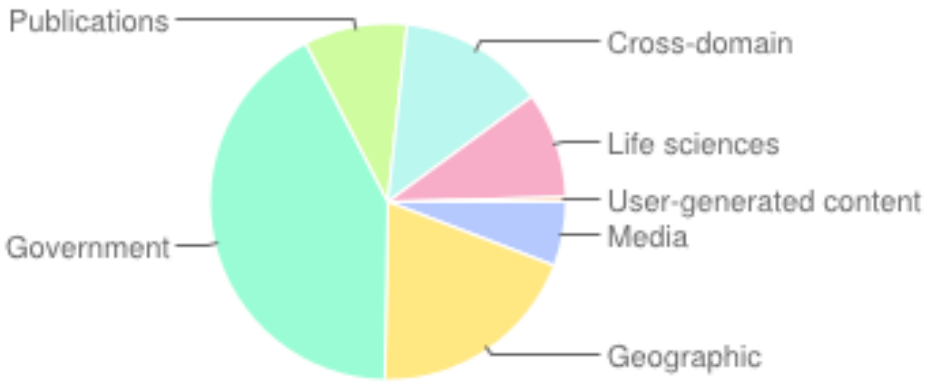
LOD today


LOV
Linked User Feedback
Slideshare 2RDF
tags2con delicious
Bricklink
Audio Scrobbler (DBTune)
Moseley Folk
GTAA
Magna-tune
DB Tropes
John Peel (DBTune)
Music Brainz (Data Incubator)
Music Brainz (zitgist)
Lotico
Klapp-stuhl-club
Semantic Tweet
yovisto
Sussex Reading Lists
St. Andrews Resource Lists
NTU Resource Lists
Man-chester Reading Lists
Open Library
Hellenic FBD
Hellenic PD
EUTC Produc-tions
Surge Radio
Discogs (Data Incubator)
Jamendo (DBtune)
RDF ohloh
Linked Crunch-base
Ontos News Portal
Source Code Ecosystem Linked Data
Plymouth Reading Lists
Open Library (Talis)
NDL subjects
t4gm info
RAMEAU SH
LEM
Crime Reports UK
Crime (En-AKTing)
business data.gov.uk
FanHubz
Last.FM artists (DBTune)
Music Brainz (DBTune)
gnoss
SSW Thesaur us
LinkedL CCN
Thesau-rus W
LCSH
LIBRIS
ntnusc
Ox Points
Popula-tion (En-AKTing)
research data.gov.uk
patents data.gov.uk
Last.FM (rdfize)
Classical (DB Tune)
Poké-pédia
Pokedex
Good-win Family
Didactal ia
flickr wrappr
my Experi-ment
theses.fr
MARC Codes List
PSH
Rådata nå!
reegle
NHS (En-AKTing)
Ren. Energy Genera-tors
Mortality (En-AKTing)
educatio n.data.gov.uk
OpenEI
BBC Program mes
BBC Music
Chronic-ling America
Event Media
Linked MDB
semantic web.org
SW Dog Food
Sudoc
IdRef Sudoc
ndlna
GND
VIAF
Norwe-gian MeSH
EEA
Energy (En-AKTing)
CO2 Emission (En-AKTing)
Ord-nance Survey
Openly Local
Recht-spraak.nl
BBC Wildlife Finder
Portu-guese DBpedia
Revyu
DDC
Calames
UB Mann-heim
Deutsche Bio-graphie
lobid Resources
Europeana
EU Insti-tutions
Open Election Data Project
legislation data.gov.uk
UK Post-codes
statistics data.gov.uk
Tele-graphis
New York Times
URI Burner
Greek DBpedia
Open Calais
RDF Book Mashup
data bnf.fr
P20
Ulm
NSZL Catalog
Swedish Open Cultural Heritage
GovWILD
Brazilian Poli-ticians
ESD stan-dards
reference data.gov.uk
data.gov.uk intervals
LOIUS
Taxon Concept
Geo Names
World Fact-book
Freebase
iServe
OS
BNB
BibBase
ECS South-ampton
ECS South-ampton EPrints
ECS (RKB Explorer)
Wiki
lobid Organi-sations
STW
GESIS
Lichfield Spen-ding
ISTAT Immi-gration
Scotland Pupils & Exams
transport data.gov.uk
NASA (Data Incu-bator)
Fishes of Texas
Geo Species
Uberblic
DBpedia
Project Guten-berg
data dcs
DBLP (FU Berlin)
DBLP (L3S)
DBLP (RKB Explorer)
Budapest
OAI
Pisa
RESEX
Scholaro-meter
Traffic Scotland
London Gazette
TWC LOGD
Eurostat
Euro-stat (FUB)
Geo Linked Data
dbpedia lite
TCM Gene DIT
data-open-ac-uk
ERA
UN/LOCODE
ACM
IRIT
IBM
NVD
Data Gov.ie
CORDIS (RKB Explorer)
CORDIS (FUB)
UMBEL
YAGO
lingvoj
Daily Med
Disea-some
SIDER
Eurécom
CiteSeer
RAE2001
New-castle
DEPLOY
LOCAH
GovTrack
Eurostat (Ontology Central)
Linked Sensor Data (Kno.e.sis)
riese
Open Cyc
Lexvo
Enipedia
LinkedCT
Drug Bank
Pfam
Roma
Course-ware
EURES
Linked EDGAR (Ontology Central)
US SEC (rdfabout)
totl.net
UniProt
PDB
VIVO Indiana
ePrints
dotAC
IEEE
RISKS
FTS
Semantic XBRL
US Census (rdfabout)
EUNIS
Twarql
WordNet (VUA)
Cornetto
Taxono my
UniProt (Bio2RDF)
PRO-SITE
ProDom
HGNC
STITCH
VIVO Cornell
LAAS
KISTI
NSF
Scotland Geo-graphy
GeoWord Net
Finnish Munici-palities
Piedmont Accomo-dations
El Viajero Tourism
SMC Journals
Climbing
Linked GeoData
WordNet (W3C)
WordNet (RKB Explorer)
LODE
SISVU
Affy-metrix
PubMed
Gene Ontology
SGD
Pub Chem
KEGG Drug
VIVO UF
JISC
ECCO-TCP
bible ontology
PBAC
Italian public schools
Ocean Drilling Codices
AEMET
Metoffice Weather Forecasts
Alpine Ski Austria
GEMET
AGROV OC
ChEMBL
OMIM
InterPro
MGI
GeneID
KEGG Pathway
KEGG Enzyme
KEGG Reaction
Turismo de Zaragoza
Janus AMP
Yahoo! Geo Planet
Weather Stations
EARTh
Product DB
Open Data Thesau-rus
Linked Open Colors
Smart Link
Medi Care
KEGG Glycan
KEGG Com-pound
National Radio-activity JP
Airports
Product Types Ontology
Italian Museums
Google Art wrapper
UniParc
UniRef
UniSTS
Chem2 Bio2RDF
Homolo Gene
Sears
Open Corpo-rates
Amster-dam Museum
medu-cator
Linked Open Numbers
Reactome
OGOLOD
UniPath way
Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2011

LOD

- 295 datasets
- >31B triples

  (Aug 2011)
- Distribution of triples



http://www4.wiwiss.fu-berlin.de/lodcloud/state/

LOD domains

Domain	Number of datasets	Triples	%	(Out-)Links	%
Media	25	1,841,852,061	5.82 %	50,440,705	10.01 %
Geographic	31	6,145,532,484	19.43 %	35,812,328	7.11 %
Government	49	13,315,009,400	42.09 %	19,343,519	3.84 %
Publications	87	2,950,720,693	9.33 %	139,925,218	27.76 %
Cross-domain	41	4,184,635,715	13.23 %	63,183,065	12.54 %
Life sciences	41	3,036,336,004	9.60 %	191,844,090	38.06 %
User-generated content	20	134,127,413	0.42 %	3,449,143	0.68 %
	295	31,634,213,770		503,998,829	

Cross-domain data

- Spanning multiple domains such as:
  - Dbpedia
  - Freebase
  - YAGO
  - UMBEL
  - OpenCyc

DBpedia

- From Wikipedia articles
  - City of Birmingham
    - http://en.wikipedia.org/wiki/Birmingham
  - DBpedia URI
    - http://dbpedia.org/resource/Birmingham
  - Wikipedia Infoboxes
    - RDF statements

Geographic Data

- Geonames, http://www.geonames.org
  - 8 M locations
- *LinkedGeoData*

Media data

- BBC
  - http://www.bbc.co.uk/programmes
  - http://www.bbc.co.uk/music
  - http://www.bbc.co.uk/wildlifefinder
  - http://news.bbc.co.uk/sport1/hi/football/world_cup_2010/default.stm
- NYTimes
  - http://data.nytimes.com/

Government data

- Data from governmental bodies and public-sector organisations
- USA, http://www.data.gov
- UK, http://data.gov.uk
- Australia, http://data.gov.au
- …

Retail and Commerce

- *GoodRelations,* http://purl.org/goodrelations
- ProductDB, http://productdb.org

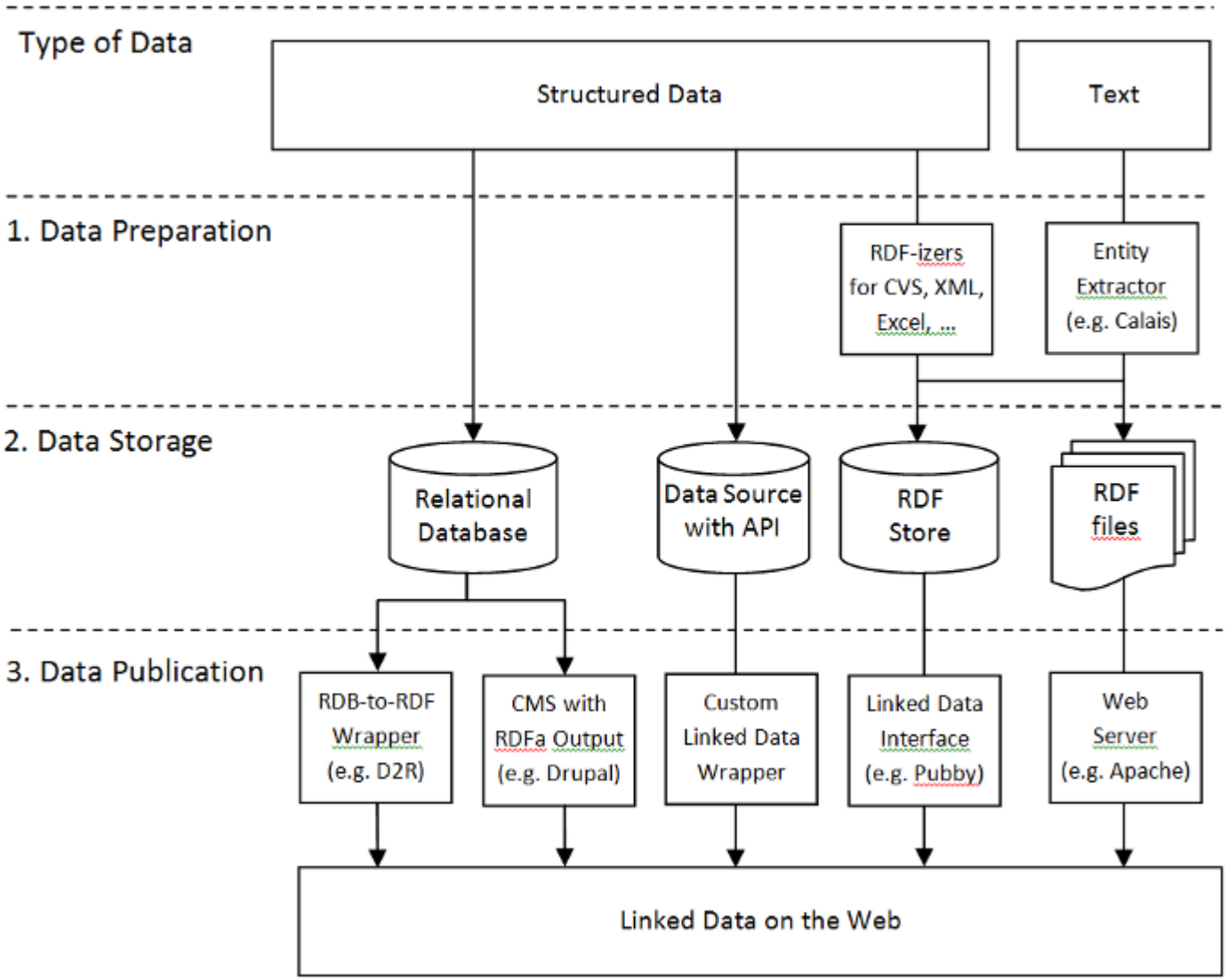
User Generated Content and Social Media

- *Semantic MediaWiki*
  - http://semantic-mediawiki.org/wiki/Semantic_MediaWiki
- *Open Graph Protocol of Facebook*
  - http://opengraphprotocol.org/

Publishing Linked Data

- Complement the existing data management infrastructures

Publishing Patterns



Patterns by input data type

- From Queryable Structured Data to Linked Data
  - Relational database to RDF wrappers
- From Static Structured Data to Linked Data
  - CSV files, Excel spreadsheets, XML files or database dumps
- From Text Documents to Linked Data
  - Parsing text using Linked Data entity extractors
    - Calais, Ontos, Dbpedia Spotlight, Nerso

Additional Considerations

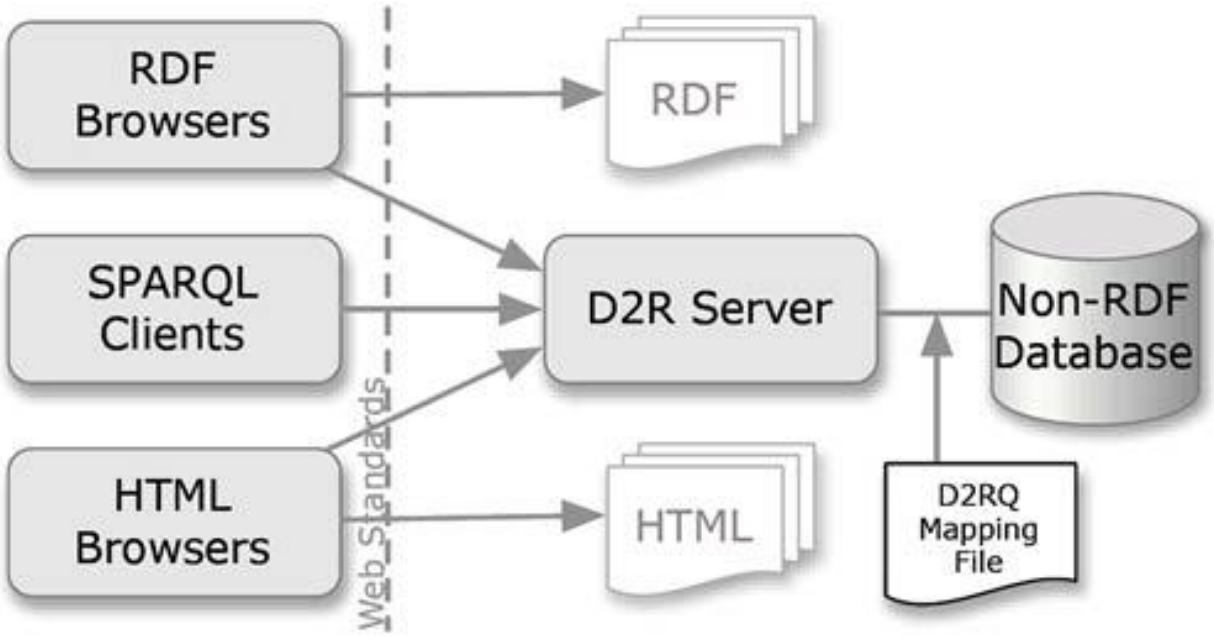
- Data Volume: How much data needs to be served?
  - Large RDF files vs. small ones vs. RDF stores
- Data Dynamism: How often does the data change?
  - Static files vs. RDF stores (dynamic)

Recipes

- Serving Linked Data as Static RDF/XML Files
- Serving Linked Data as RDF Embedded in HTML Files
  - RDFa
- Serving RDF and HTML with Custom Server-Side Scripts
- Serving Linked Data from Relational Databases
- Serving Linked Data from RDF Triple Stores
- Serving Linked Data by Wrapping Existing Application or Web APIs

Linked Data from Relational DB

- Keeping legacy data intact



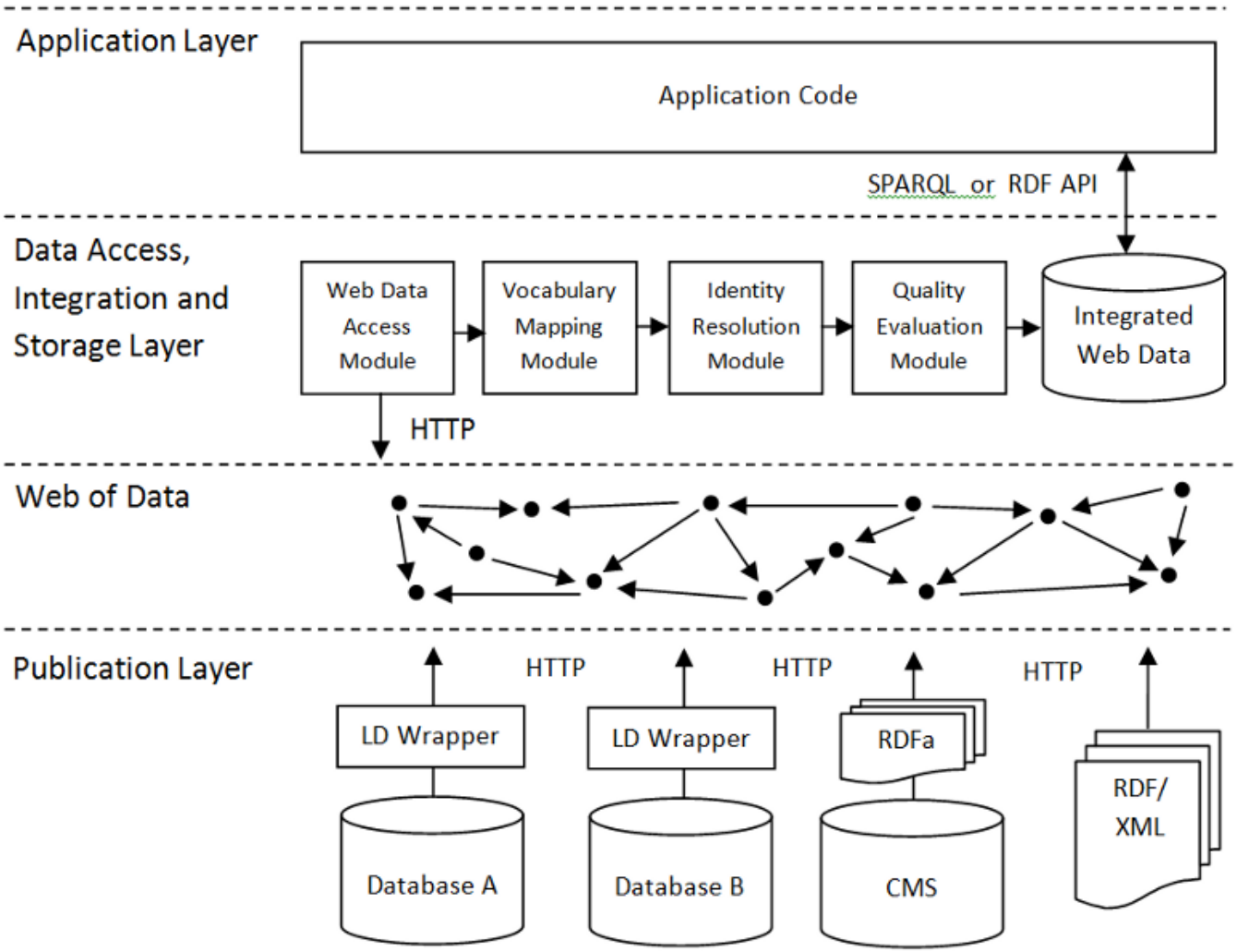
Consuming linked data

- Deployed applications
  - Generic apps
    - Linked Data Browsers
      - Disco, Tablutor, …
    - Linked Data Search Engines
      - *Sig.ma, Falcons, SWSE*
  - Domain-specific apps
- Linked data mashups

Architecture of Linked Data Apps

- Patterns
  - **Crawling Pattern**
    - Replicate data in local store, data is not fresh
  - **On-The-Fly Dereferencing Pattern**
    - Data fresh, but processing is slow
  - **Query Federation Pattern**
    - Querying multiples SPARQL endpoints
    - Performance problem

Crawling pattern



Data Access, Integration and Storage Layer

- Accessing the Web of Data. (endpoints)
- Vocabulary Mapping. (owl:equivalentClass, etc.)
- Identity Resolution. (owl:sameAs)
- Provenance Tracking. (go back to orig. resource)
- Data Quality Assessment.
- Using the Data in the Application Context.

End