Lecture Notes

Big Data in Medical Informatics

Week 8:
Sharing Biomedical Data via Semantic Web

RWTH AACHEN UNIVERSITY

Research Data Repositories and Directories

NIF NEUROSCIENCE INFORMATION FRAMEWORK

FLOW Repository

GEO Gene Expression Omnibus

DataShare Open data for the global research community

EMBL-EBI PRIDE

ALLEN BRAIN ATLAS DATA PORTAL

dbGaP

SCIENTIFIC DATA

Home | About | For Authors | Advisory and Editorial Board | Open Access | Contacts | FAQ

Helping you publish, discover, and reuse research data

Credit — Credit, through a citable publication, for depositing & sharing your data

Reuse — Complete, curated & standardized descriptions enable the reuse of your data

Quality — Rigorous community based peer review

Discovery — Find datasets relevant to your research

Open — Promotes & endorses open science principles & available to all through a Creative Commons license

Service — In-house curation, rapid peer review & publication of your data descriptions

EMBL-EBI ArrayExpress

DRYAD

Directories

Databib Find Repositories | Submit | Connect

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

Ref: Open Genomic Data Repositories and Analysis Resources , 4-26-16, Megan Laurance, Ph.D.

Why Share Data?

Sometimes you have to…

Funders

NIH National Institutes of Health
Turning Discovery Into Health

American Heart Association

Publishers

PLOS

npg nature publishing group

Cell PRESS

PNAS

Data repositories are useful when you need to comply with Funder or Publisher Policy

Ref: Open Genomic Data Repositories and Analysis Resources , 4-26-16, Megan Laurance, Ph.D.

Example of Funder Requirements for Data Sharing

- NIH Genomic Data Sharing Policy
  - Updated August, 2014 https://gds.nih.gov/pdf/NIH_GDS_Policy_Overview.pdf
  - Strikes a balance between encouraging data sharing as broadly as possible and addressing concerns re: identification of patient donors from genomic data
  - NIH database of human genotypes and phenotypes, dbGaP, will remain the required data repository for all NIH-funded human genetic studies, GEO for gene expression studies.
  - Data release is mandated at the time of publication of results or earlier
- This policy applies to NIH intramural research projects generating genomic data on or after January 25, 2014

Ref: Open Genomic Data Repositories and Analysis Resources , 4-26-16, Megan Laurance, Ph.D.

Example of Publisher Requirements for Data Sharing

- From Nature Publishing:
  - A condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to others without undue qualifications.
  - **Data sets must be made freely available to readers from the date of publication, and must be provided to editors and peer-reviewers at submission, for the purposes of evaluating the manuscript.**
  - For the following types of data set, **submission to a community-endorsed, public repository is mandatory**. Accession numbers must be provided in the paper. Examples of appropriate public repositories are listed below.
    - Microarray data
      - MIAME-compliant microarray data: deposit in GEO or ArrayExpress upon submission to the journal.
      - Data must be MIAME-compliant, as described at the MGED web site specifying microarray standards.
    - Simple genetic polymorphisms should be submitted to dbSNP.
    - For data linking genotyping and phenotyping information, we strongly recommend submission to dbGAP or EGA, two repositories that have mechanisms for access control for human health-related phenotypes.

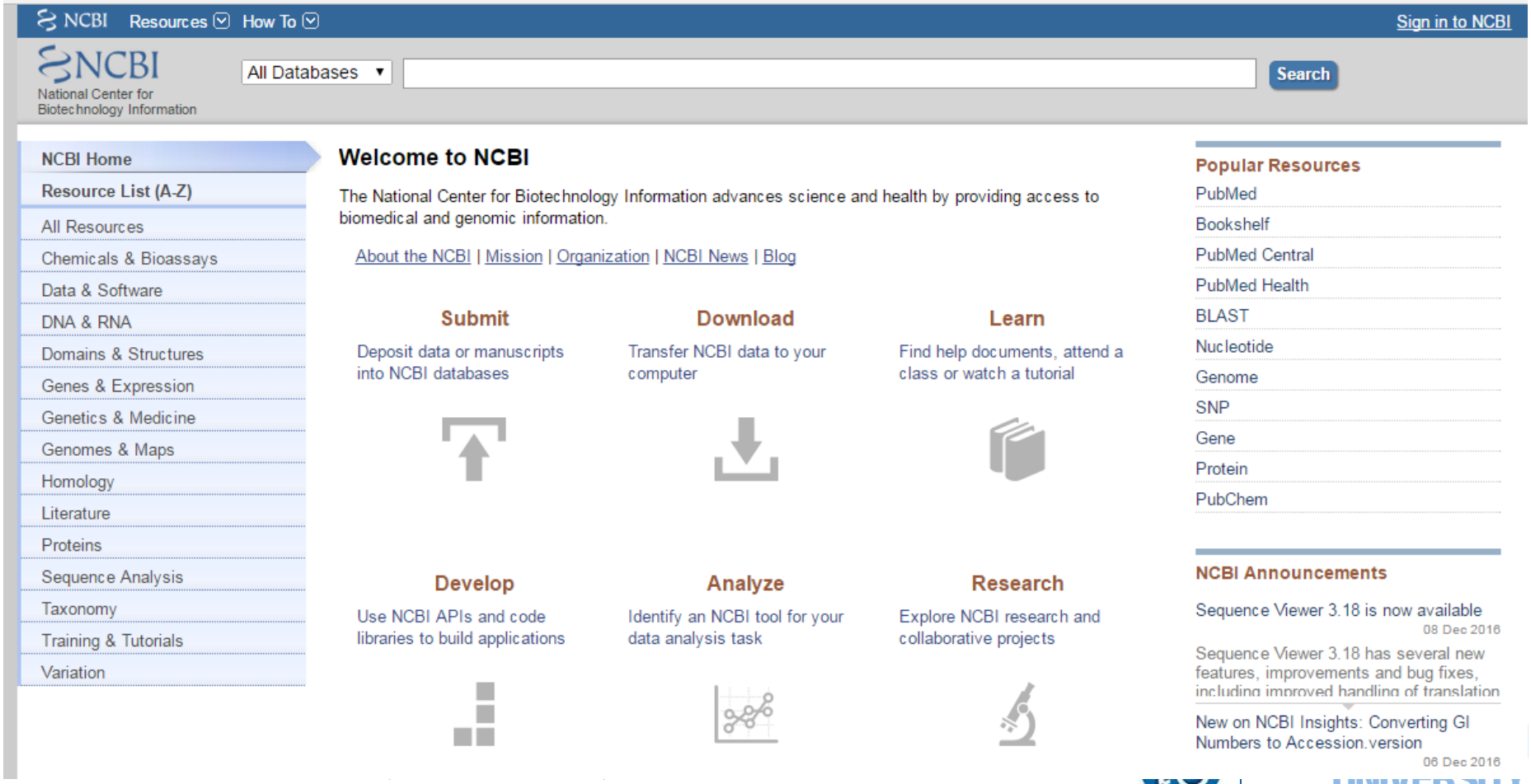
Ref: Open Genomic Data Repositories and Analysis Resources , 4-26-16, Megan Laurance, Ph.D.

Benefits of Data Sharing and Reuse

- Compare results from related published experiments to your own. Gain confidence in novel insights.
- Identify novel drug targets, fish for new genes/biomarkers associated with disease, drug response, phenotype of interest.
- Get your feet wet in an experimental method before doing it yourself.
- Develop new analysis methods and test software
- Supports more efficient use of funding by avoiding duplicate data collection
- Encourages reproducibility/Discourages fraud

NCBI

- The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health.
- https://www.ncbi.nlm.nih.gov/



https://www.ncbi.nlm.nih.gov/gap/phegeni

NCBI Resources How To

PheGenI
Phenotype-Genotype Integrator

All Databases

Search Summary

Search Criteria
Phenotype Selection
Trait: Breast Neoplasms

Modify Search

Search Results

Association Results	1 - 50 of 222	Searched by phenotype trait.
Genes	1 - 31 of 31	Searched by gene IDs retrieved from page 1 of association results.
SNPs	1 - 24 of 24	Searched by SNP rs numbers retrieved from page 1 of association results.
eQTL Data	No data found.	Searched by SNP rs numbers retrieved from page 1 of association results.
dbGaP Studies	1 - 36 of 36	Searched by traits retrieved from page 1 of association results.
Genome View	24 SNPs and 31 genes over 13 chromosomes.	

Modify Search Show All Hide All

Search Criteria

Association Results

1 - 50 of 222 < Previous Next > Page 1 Go Download Modify Search

#	Trait	rs #	Context	Gene	Location	P-value	Source	Study	PubMed
1	Breast Neoplasms	rs2981582	intron	FGFR2	10: 123,352,317	2.000×10^{-76}	NHGRI		17529967
2	Breast Neoplasms	rs3803662	intergenic	TOX3, CHD9	16: 52,586,341	1.000×10^{-36}	NHGRI		17529967
3	Breast Neoplasms	rs2981579	intron	FGFR2	10: 123,337,335	4.000×10^{-31}	NHGRI		20453838
4	Breast Neoplasms	rs1219648	intron	FGFR2	10: 123,346,190	1.000×10^{-30}	NHGRI		21263130
5	Breast Neoplasms	rs4784227	intergenic	TOX3, CHD9	16: 52,599,188	1.000×10^{-28}	NHGRI		20585626
6	Breast Neoplasms	rs889312	intergenic	RPL26P19, MAP3K1	5: 56,031,884	7.000×10^{-20}	NHGRI		17529967
7	Breast Neoplasms	rs3803662	intergenic	TOX3, CHD9	16: 52,586,341	6.000×10^{-19}	NHGRI		17529974
8	Breast Neoplasms	rs2046210	intergenic	C6orf97, ESR1	6: 151,948,366	2.000×10^{-15}	NHGRI		19219042
9	Breast Neoplasms	rs614367	intergenic	IFITM9P, CCND1	11: 69,328,764	3.000×10^{-15}	NHGRI		20453838
10	Breast Neoplasms	rs3803662	intergenic	TOX3, CHD9	16: 52,586,341	3.000×10^{-15}	NHGRI		20453838
11	Breast Neoplasms	rs10488592	intergenic	SEMA3A, RPL7P30	7: 83,944,353	3.432×10^{-15}	dbGaP	phs000007	17903305

NIH Cancer Genome Atlas TCGA

The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI)

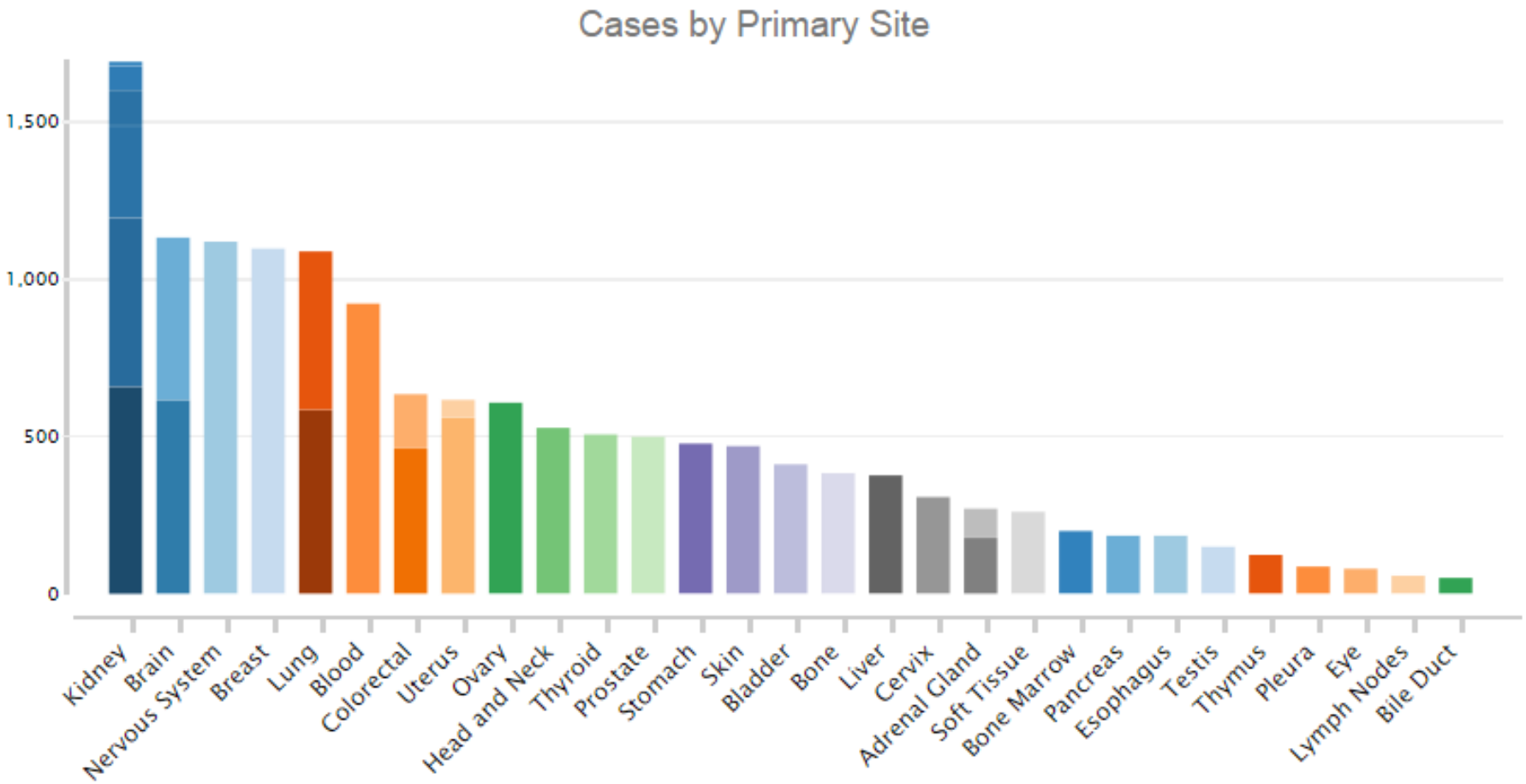
- The project scheduled 500 patient samples, more than most genomics studies, and used different techniques to analyze the patient samples.
  - gene expression profiling,
  - copy number variation profiling,
  - SNP genotyping,
  - genome wide DNA methylation profiling,
  - microRNA profiling, exon sequencing.

- publically available
- has been used widely by the research community.

- https://cancergenome.nih.gov/

NIH Cancer Genome Atlas TCGA

- 33 types of cancer.
- 2.5 petabytes of data describing tumor tissue and matched normal tissues from more than 11,000 patients,



Re-analysis of TCGA gene expression and copy number alteration data

Cell Death and Survival

Molecular Cancer Research

MTBP Is Overexpressed in Triple-Negative Breast Cancer and Contributes to Its Growth and Survival

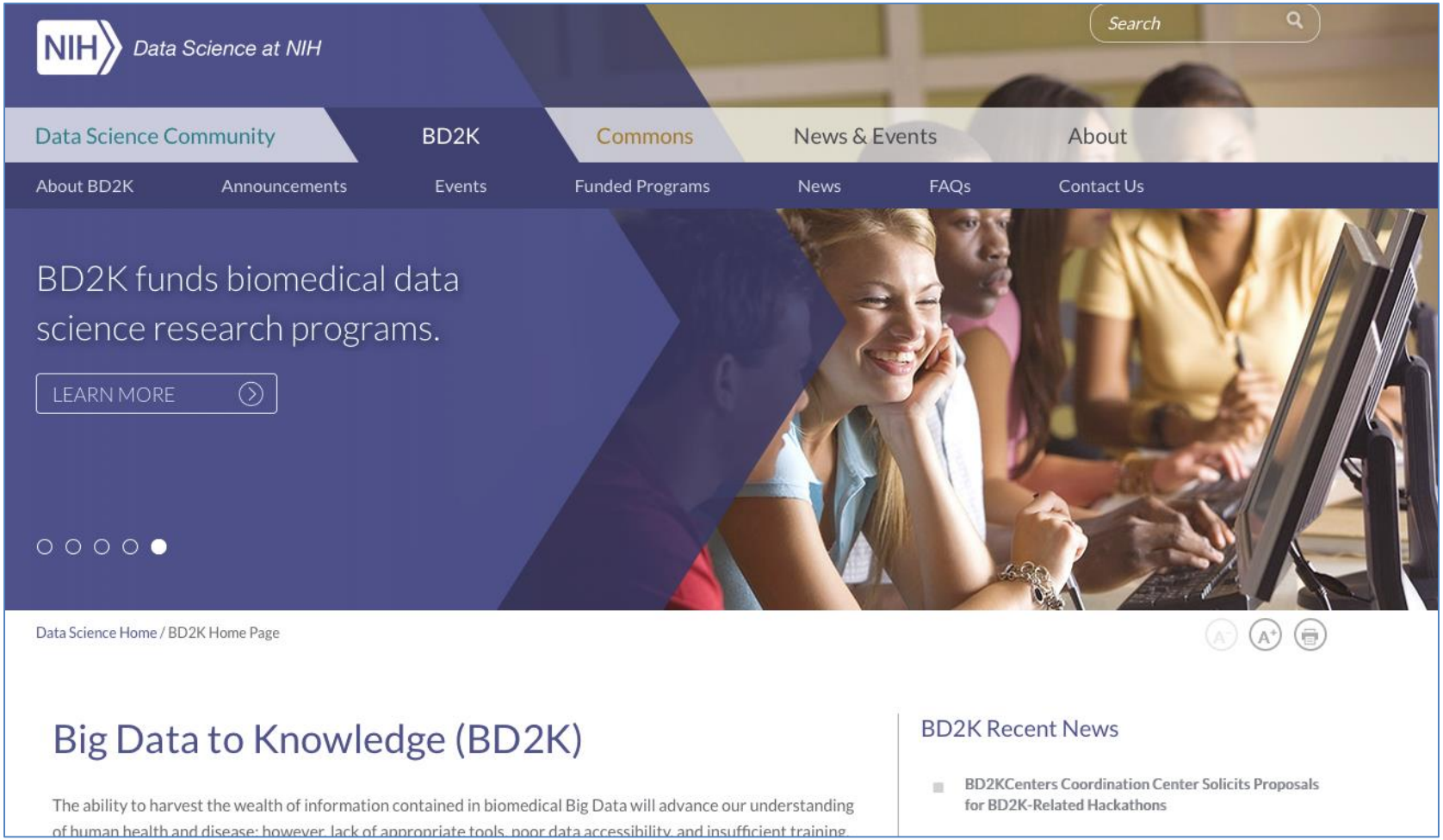
Brian C. Grieb[1], Xi Chen[2], and Christine M. Eischen[1]

http://www.ncbi.nlm.nih.gov/pubmed/24866769

"By using TCGA data, researchers found that MTBP is expressed at different levels in TNBC subtypes. These findings may have positive implications for further study and future treatments. The data used in this research, and all available TCGA data, can be found through the TCGA Data Portal and the cBioPortal."

http://cancergenome.nih.gov/

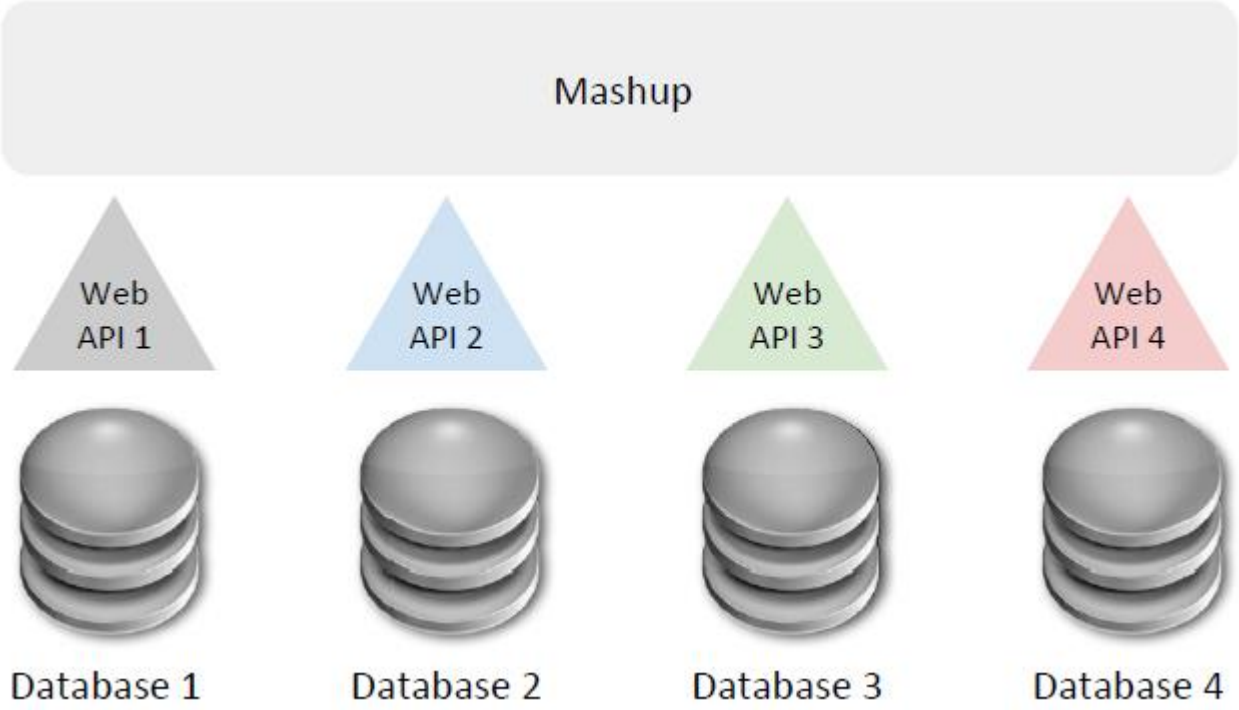
NIH Big Data 2 Knowledge Initiative



Data Sharing in Biomedical Domain

- Biomedical portals servers variety of data sets.
- Researcher can access, download and integrate them
- There is a number of different (proprietary) **Web APIs**, data exchange formats, and Mashups on top of that.

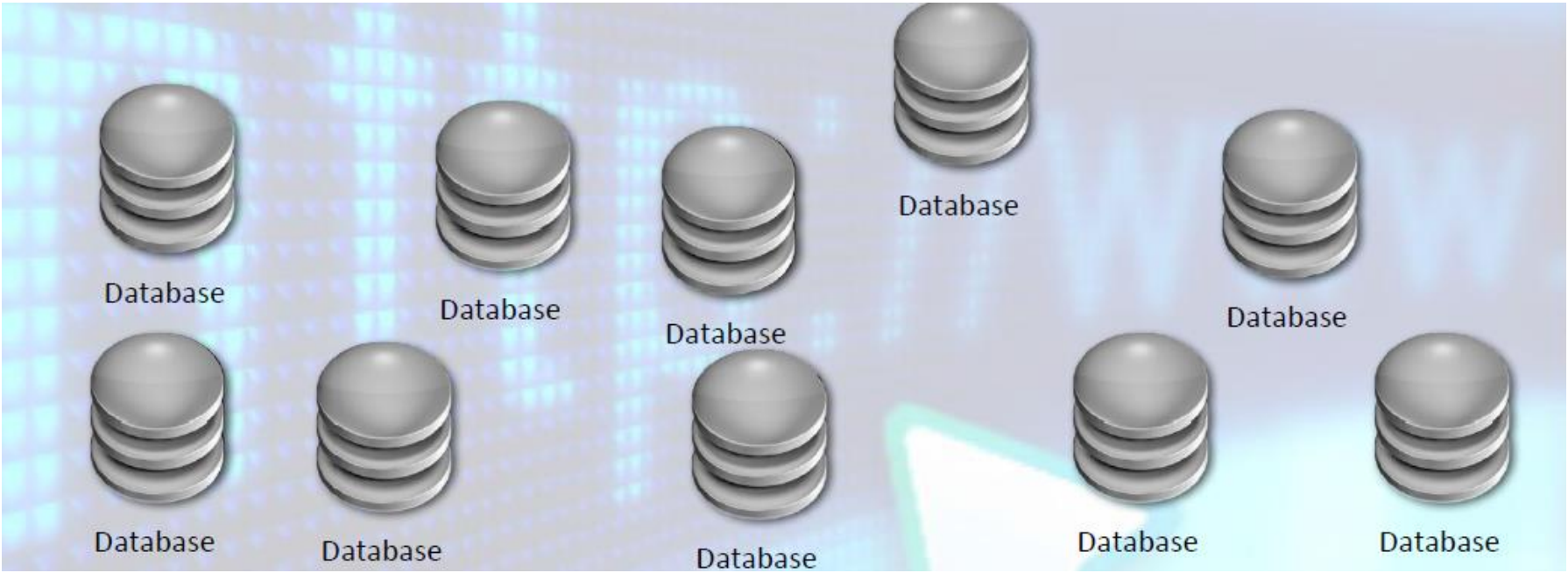
Question: Is it the most effective way of data sharing ?



Data Sharing in Biomedical Domain

What do we need more ?

- Data is locked up in small data islands



A Network of Data and Knowledge

- Interconnected
- Universal
- All encompassing

- assists humans, organisations and systems with problem solving
- enabling innovation and increased productivity

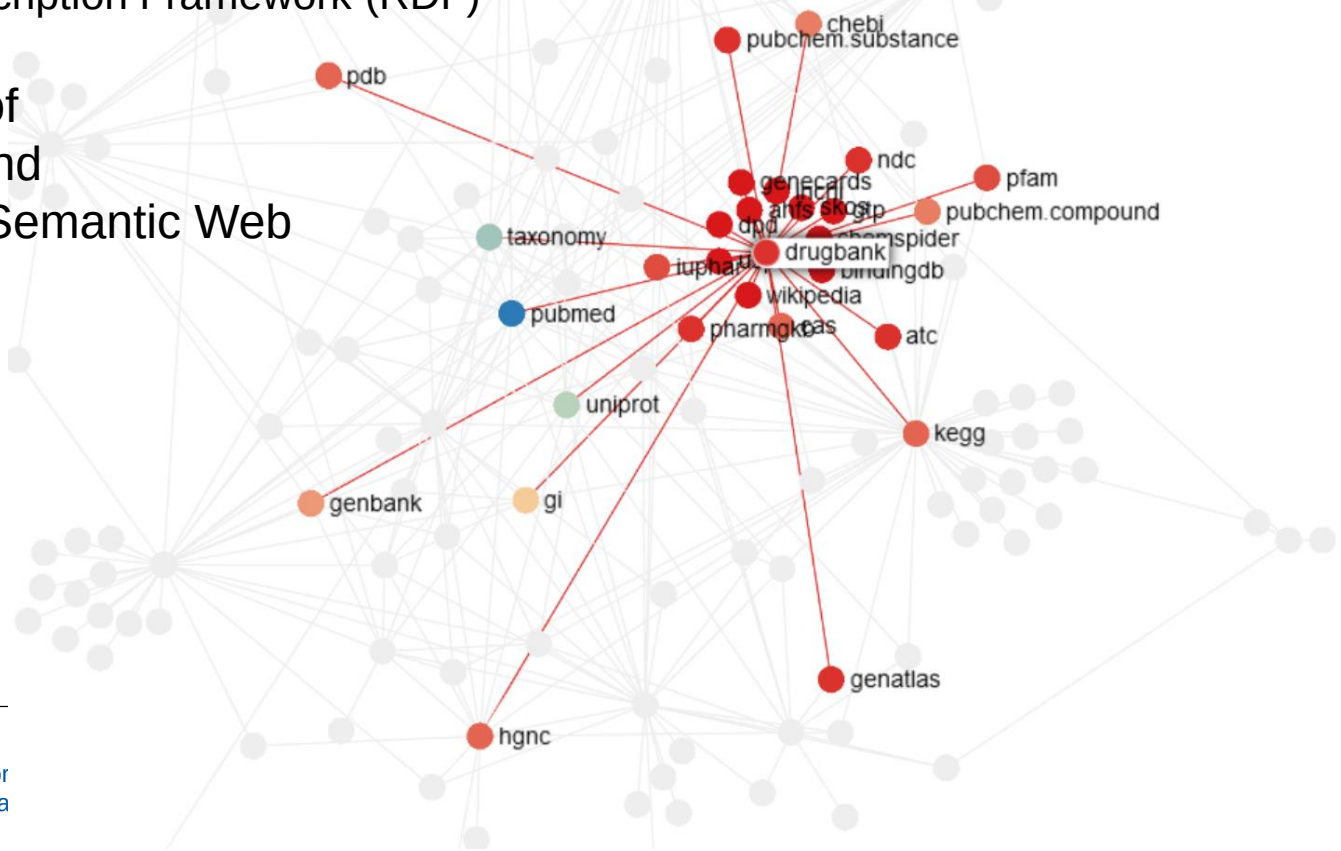
Linked Data

- Entities (people, proteins, pathways, etc) are identified using globally unique identifiers (URIs)
- Entity descriptions are represented with a standardized language (RDF)
- Data can be retrieved using a universal protocol (HTTP)
- Entities (concepts, data, resources) can be linked together to increase interoperability

Publishing Biomedical Data as Linked Data

- *Linked Data* is an approach to data integration that employs
  - ontologies,
  - terminologies,
  - Uniform Resource Identifiers (URIs) and
  - the Resource Description Framework (RDF)

- to connect pieces of
- data, information and
- knowledge on the Semantic Web



Publishing Biomedical Data as Linked Data

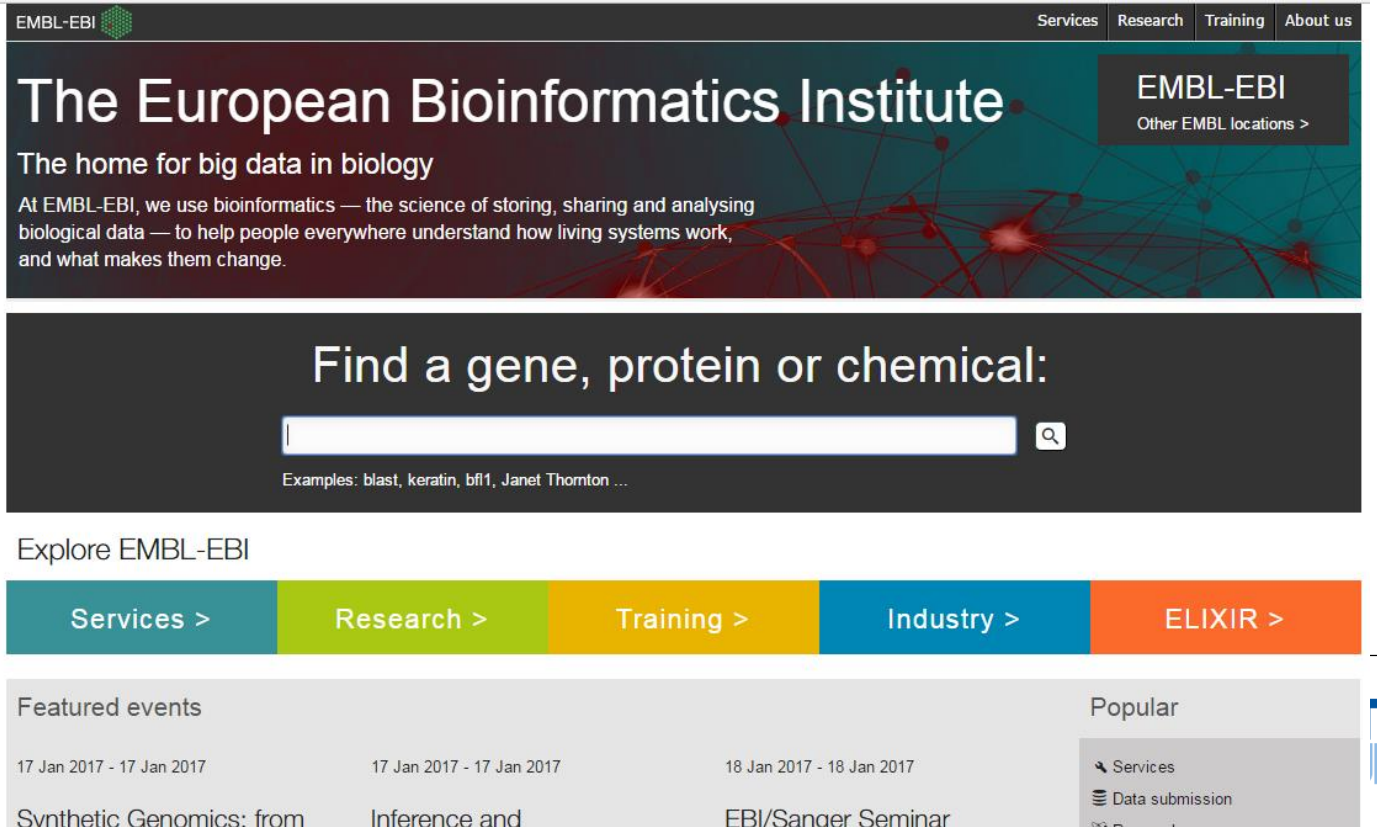
- Possible motivations to publish Linked Data sets :
- Shareability: A data provider or publisher would like to make some existing data more openly accessible, through standard, programmatic interfaces such as SPARQL or resolvable URIs.
- Integration: A developer desires to create and maintain a list of links between different RDF data sets so that she can easily query across these datasets.

Publishing Biomedical Data as Linked Data

- Semantic Normalization: A computer science researcher is interested in indexing an existing RDF data set using a set of common ontologies, so that the dataset can be queried using ontological terms.
- Discoverability: A bench biologist would like to be able to discover what is available in the Semantic Web about a set of proteins, genes or chemical components, either as published results, raw data, or tissue libraries.
- Federation: A pharmaceutical company desires to retrieve data from sources distributed across its enterprise using SPARQL.

The EBI RDF platform: linked open data for the life sciences

- The European Bioinformatics Institute (EBI) is the largest bioinformatics resource provider in Europe.
- EBI databases are accessible via dedicated interfaces, web services, data download and (in a few cases) direct database access.
- The EBI RDF platform has been developed to meet an increasing demand to coordinate RDF activities across the institute and provides a new entry point to querying and exploring integrated resources available at the EBI.



EMBL-EBI

RDF Platform

RDF Platform | Services | Documentation | About | FAQ

The *EBI RDF Platform* aims to bring together the efforts of a number of EMBL-EBI resources that provide access to their data using Semantic Web technologies. It provides a unified way to query across resources using the W3C SPARQL query language. We welcome **comments or questions** via our feedback form.

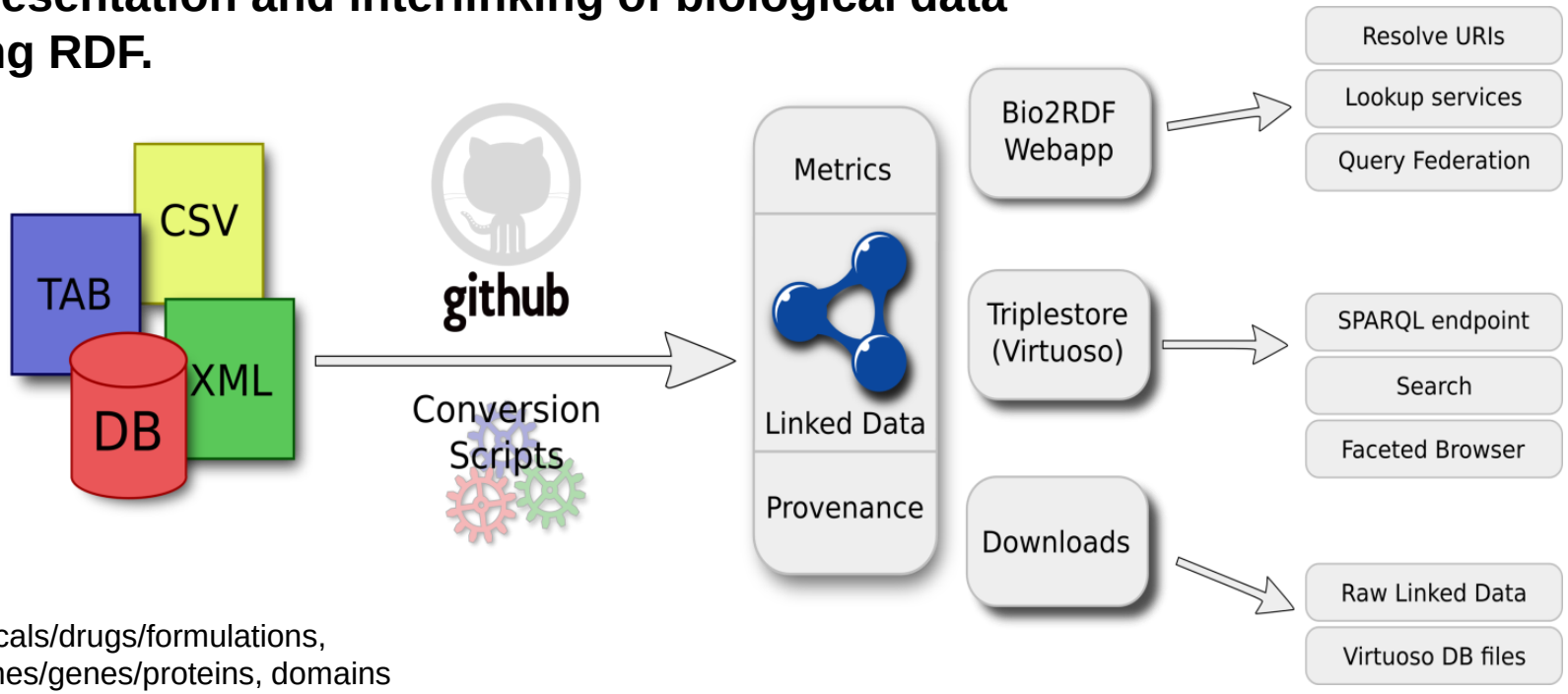
https://www.ebi.ac.uk/rdf/

Current RDF resources

Services	Quick links	Example query
BioModels	Service description SPARQL endpoint Documentation RDF download	All model elements with annotations to acetylcholine-gated channel complex (GO:0005892)
BioSamples	Service description SPARQL endpoint Documentation RDF download	Samples treated with alcohol
ChEMBL	Service description SPARQL endpoint Documentation RDF download	Find drug-like (but currently not approved) molecules which bind 7TM1 GPCRs with high affinity
Ensembl	SPARQL endpoint Documentation RDF download	Get all the genes, transcripts and exons on a chromosome
Expression Atlas	Service description SPARQL endpoint Documentation RDF download	Under what experimental conditions is Ensembl gene ENSG00000129991 (TNNI3) expressed?

RWTH AACHEN UNIVERSITY

Bio2RDF is an open source project to unify the representation and interlinking of biological data using RDF.



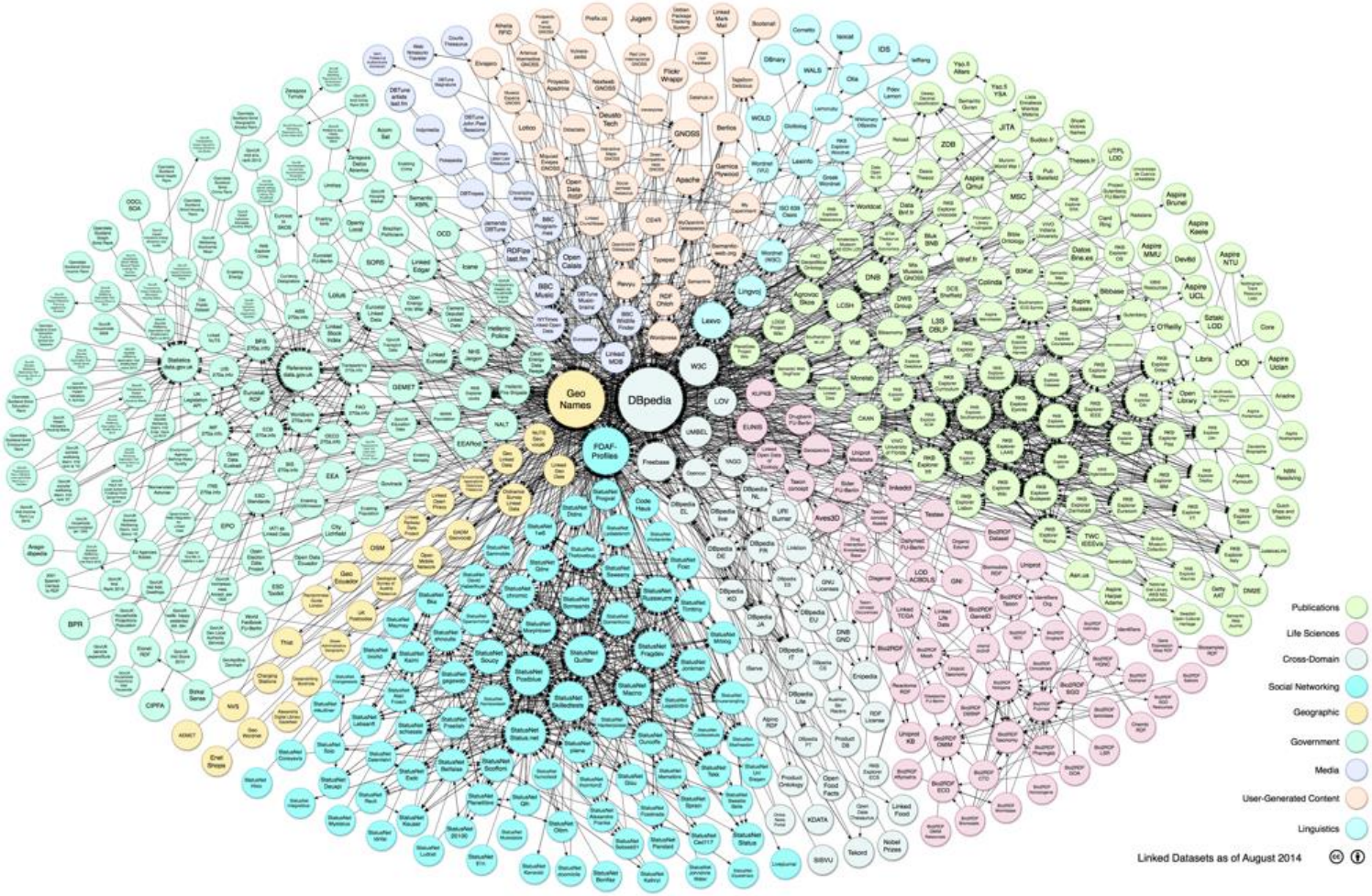
chemicals/drugs/formulations,
genomes/genes/proteins, domains
Interactions, complexes & pathways
animal models and phenotypes
Disease, genetic markers, treatments
Terminologies & publications

- 11B+ interlinked statements from 35 biomedical datasets and 400+ ontologies
- dataset description, provenance & statistics
- **A growing interoperable ecosystem with the EBI, NCBI, DBCLS, NCBO, OpenPHACTS, and commercial tool providers**

Bio2RDF
normalizes identifiers, formats, links, and access

LOD Cloud

- *Linked Open Data* : a data cloud, whose resources are published on the web using the *Linked Data* technology-
- Pink ones are Life Sciences data sets



RWTH AACHEN UNIVERSITY

Semantic Web

The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation

Semantic Web

- For **communication**,
  - information has to be correctly transmitted (**Syntax**)
  - the meaning (**Semantics**) of the transmitted information
  - must be interpreted correctly (= **understanding**)
- **Understanding** depends on
  - the **context** of both sender and receiver and
  - the **pragmatics** of the sender
- **Context** of sender and receiver depend on
  - the experience (knowledge of the world) of both sender and receiver

Web Standards

1. RDF – Resource Description Framework

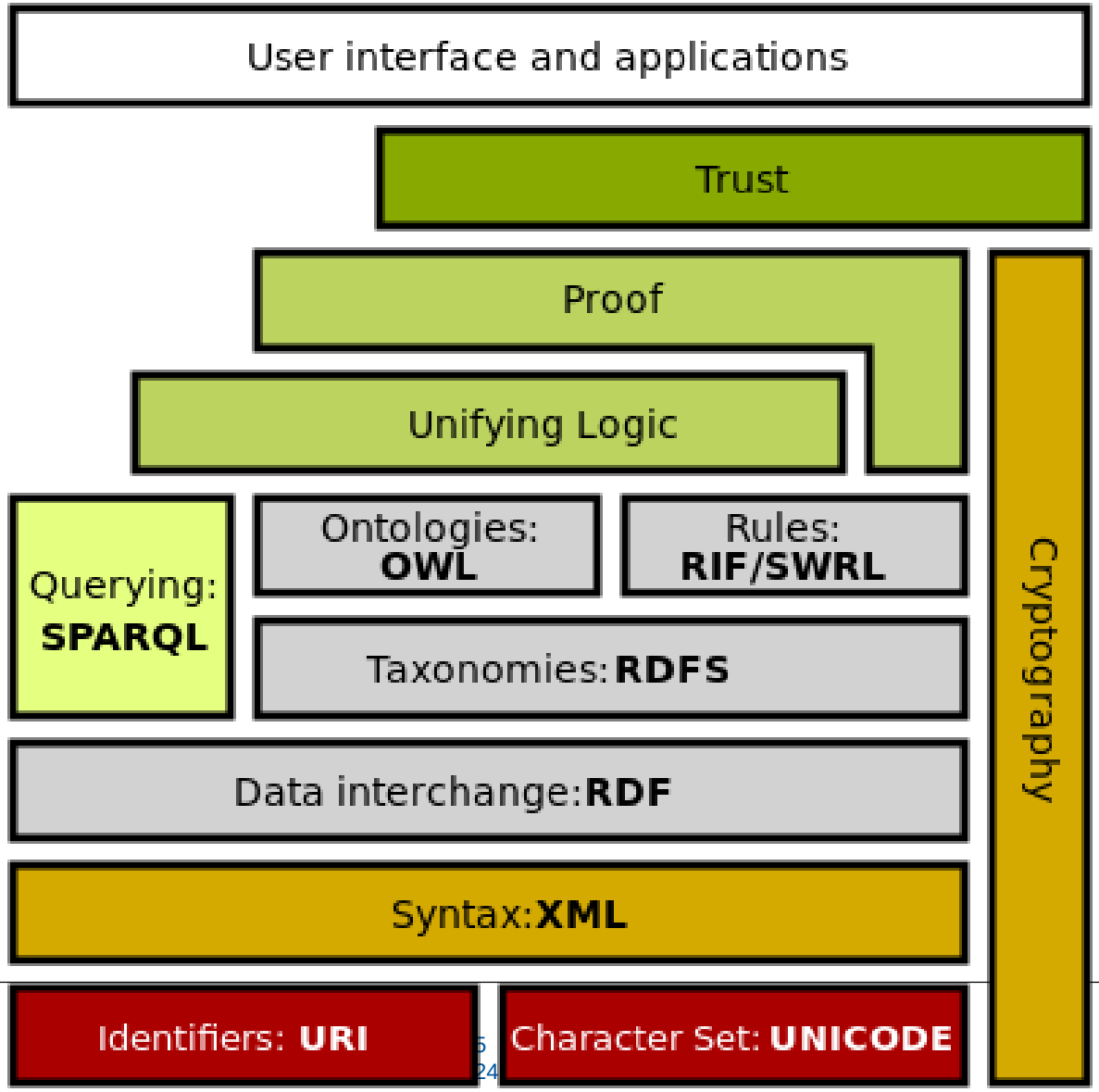
   Graph based Data – nodes and arcs

   – Identifies objects (URIs)
   – Interlink information (Relationships)

2. Vocabularies (Ontologies)

   – provide shared understanding of a domain
   – organise knowledge in a machine-comprehensible way
   – give an exploitable meaning to the data

Semantic Web Stack



RDF

- RDF = Resource Description Framework
  - W3C Recommendation since 1998
    - http://www.w3.org/RDF
  - Version 1.1 since 2014
    - http://www.w3.org/TR/rdf11-concepts/

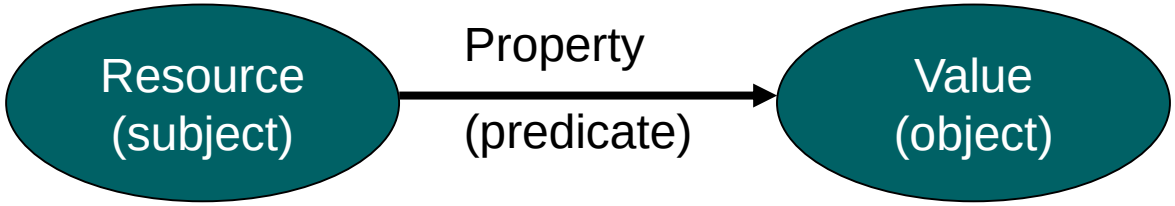
- RDF is a data model
  - Originally used for metadata for web resources, then generalized
  - Encodes structured information
  - Universal, machine readable exchange format

- Data structured in graphs
  - Vertices, edges

RDF

- Resource
  - can be everything
  - must be uniquely identified and referenceable via URI
- Description
  - = description of resources
  - via representing properties and relationships among resources as graphs
- Framework
  - = combination of web based protocols (URI, HTTP, XML, Turtle, JSON, …)
  - based on formal model (semantics)

- Knowledge in RDF is expressed as a list of statements
- all RDF statements follow the same simple schema (= RDF Triple)



Subject has a ***property*** with value "***object*** " (s,p,o)

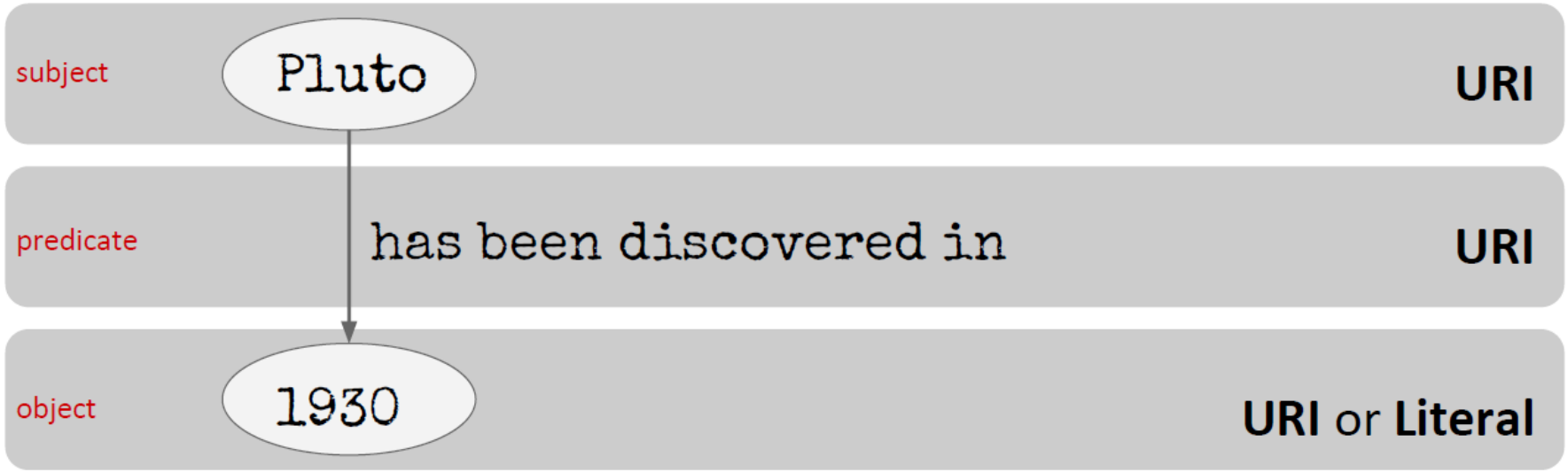
Basic Ideas behind RDF

RDF uses Web identifiers (URIs) to identify resources

RDF builds relationships between resources

RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

RDF



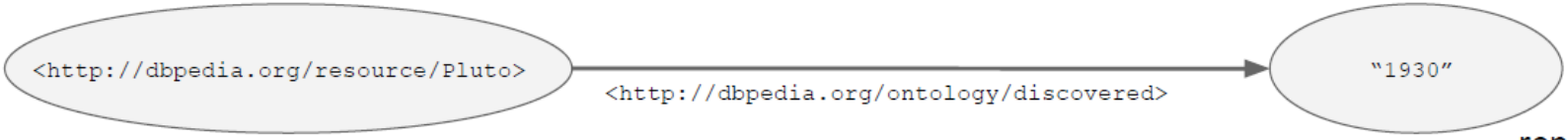
Ref: Linked Data Engineering , Prof. Dr. Harald Sack, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure & Karlsruhe Institute of Technology

- **RDF Statements (RDF-Triple):**

Subject	+	Property	+	Object / Value	
URI		**URI**		**URI / Literal**	RDF Building Blocks

N-Triples Serialization

```
<http://dbpedia.org/resource/Pluto>    <http://dbpedia.org/ontology/discovered>    "1930" .
```



graph representation

Ref: Linked Data Engineering , Prof. Dr. Harald Sack, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure & Karlsruhe Institute of Technology

Triple

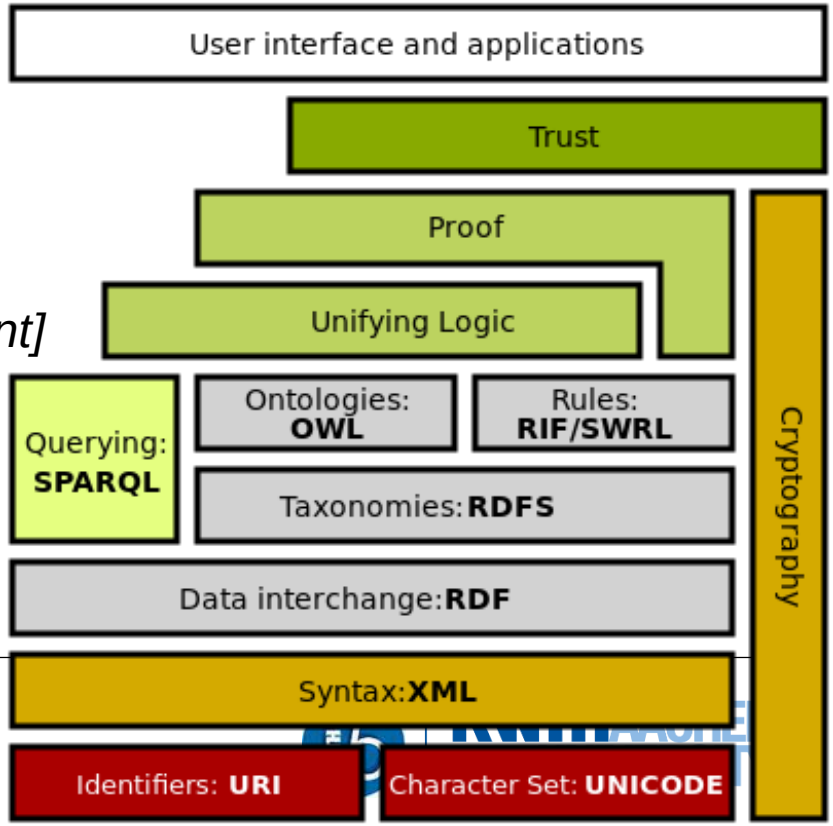
- A Resource (Subject) is anything that can have a URI: URIs or blank nodes
- A Property (Predicate) is one of the features of the Resource: URIs
- A Property value (Object) is the value of a Property, which can be literal or another resource: URIs, literal, blank nodes

URIs and Unicode

- URI = Uniform Resource Identifier
  - Used to create globally unique names for resources
  - Every object with clear identity can be a resource
  - Books, places, organizations ...
  - In the books domain the ISBN serves the same purpose

- IRIs: Unicode-aware extension of URIs (I = Internationalized)
- See RFC 3987: https://tools.ietf.org/html/rfc3987

- Typically hierarchical structure
  - *[scheme:][//authority][path][?query][#fragment]*



What are URIs?

- **URI** = Uniform Resource Identifier
- Used for worldwide, unique identification of resources
- Every object (in the context of the application) maybe a resource
  - As long as it has a unique identity
  - E.g. books, places, people, relation between those things, abstract concepts
- Unique Identifiers were already used for other and more specific domains, e.g. ISBN for books or tax identification numbers for people
- Extension of the URL concept:
  - Not every URI belongs to a webpage, *but* often a URL is used as a URI for web pa

Syntax of URIs

- Tim Berners-Lee submitted 1994 the RFC 1630 about URIs
  - http://www.ietf.org/rfc/rfc1630.txt (current version: RFC 3986 of 2009)
  - Starts with the URI schema
  - Protocol (e.g. http, ftp, mailto) and hierarchy separated by ':'
    - Queries parameters can be appended using a leading '?'
    - Fragment identifiers can be appended using a leading '#'
- protocol ":" hierarchy ["?" query] ["#" fragment]

```
http://en.wikipedia.org/w/index.php?search=rdf
http://en.wikipedia.org/wiki/Resource_Description_Framework#Examples
```

Fragment Identifier

- Fragment identifier is a short string of characters that refers to a resource that is subordinate to another, primary resource.
- The primary resource is identified by a Uniform Resource Identifier (URI), and the fragment identifier points to the subordinate resource.
- … with the help of URI references (with “#”-attached fragments) or content negotiation

- Example: URI for Shakespeare's "Othello":
  - bad (why?): *http://de.wikipedia.org/wiki/Othello*
  - good: *http://de.wikipedia.org/wiki/Othello#URI*

Self-defined URIs

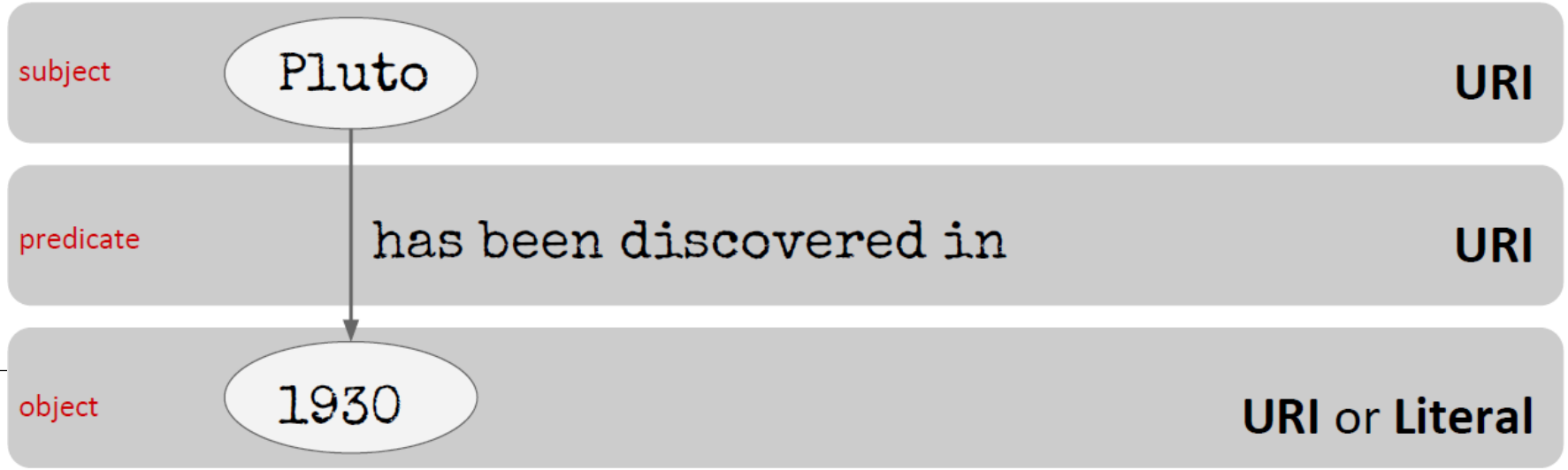
- Needed if a resource has no URI yet
- Possible strategy to avoid overlapping URIs
  - Use HTTP URIs of your own webspace!
  - It is also possible to publish documentation of the URI at this place
  - E.g. http://jens-lehmann.org/foaf.rdf#i (a person, not a document)

Other Identification Systems

- IRI = **Internationalized Resource Identifier**
  - Generalization of URI, can contain Unicode characters
  - E.g. *http://www.example.org/Wüste*
- URN = **Uniform Resource Name**
  - Subset of URIs, used for identifying resource with freely choosable names
  - Intended for worldwide unique and persistent identification
  - E.g. *urn:issn:0167-6423* URN of a Spider Man movie
- ISBN = **International Standard Book Number**
  - E.g. *ISBN 978-3-86680-192-9*
- ISSN = **International Standard Serial Number**
  - E.g. *ISSN 1234-5678*
- DOI = **Digital Object Identifier**
  - E.g. *DOI 10.1000/182*

Literals

- Used to model data values
- Representation as strings
- Interpretation through datatype
- Literals may **never be the origin of a node** of an RDF graph
- Edges may **never be labeled** with literals
- Language Declaration: Two letter language modifier: “Aachen"@de



Literals and Data Types

- Typed literals can be expressed via XML Schema datatypes
- Namespace for typed literals:
  http://www.w3.org/2001/XMLSchema#

- Examples:

  "Semantics"^^<http://www.w3.org/2001/XMLSchema#string>
  “ 1161.00"^^<http://www.w3.org/2001/XMLSchema#float>
  "2015-08-02"^^<http://www.w3.org/2001/XMLSchema#date>

- Language Tags denote the (natural) language of the text:
  Example:
  "Semantik"@de , "Semantics"@en

Data Types

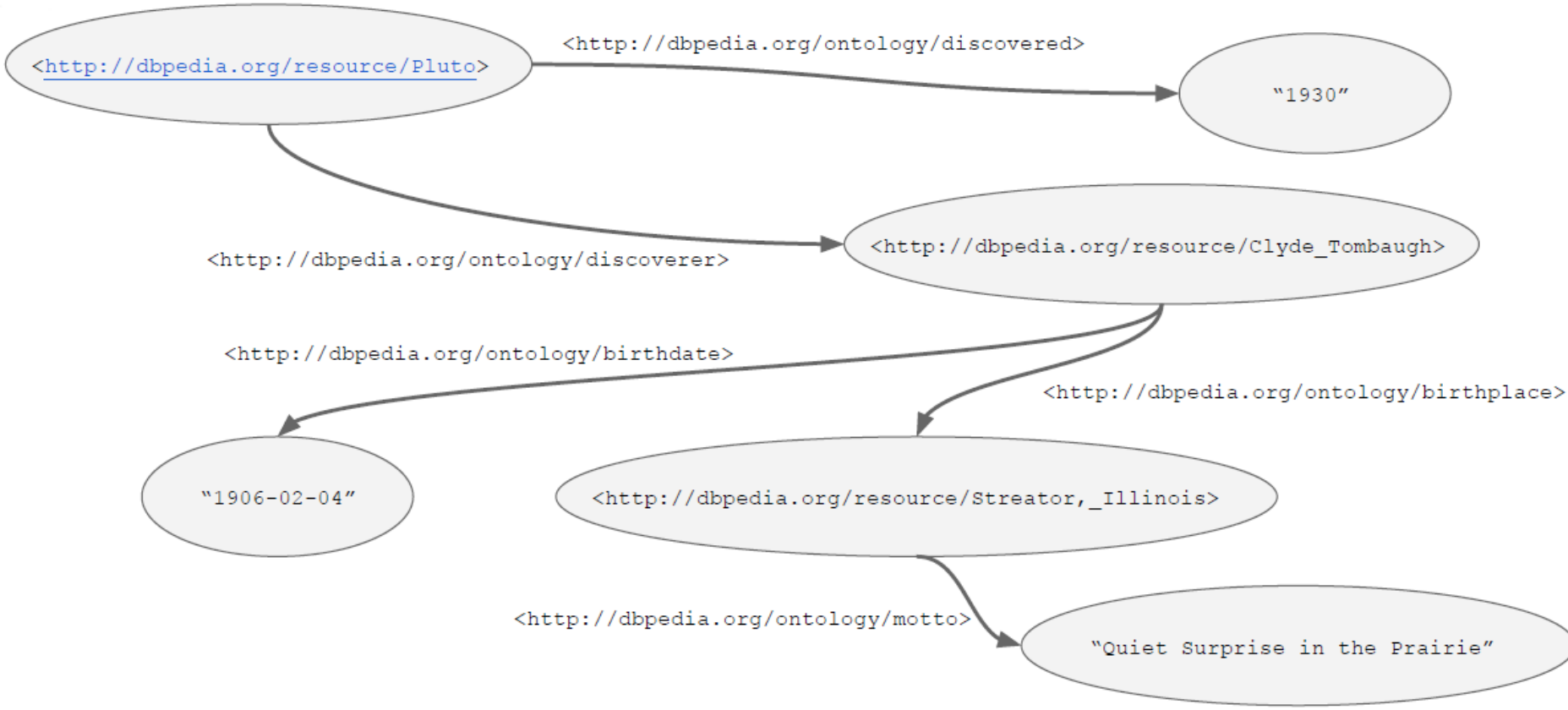
- Untyped literals are treated as strings: *"02" < "100" < "11" < "2"*
- Datatypes get identified by URIs and are freely choosable
- Typically usage of XML Schema Datatypes (XSD)
- Syntax: *"data value"^^<datatype-URI>*
- *rdf:HTML* and *rdf:XMLLiteral* are the only predefined datatypes in RDF
  - Used for HTML and XML fragments
- Example:
  - "123“^^http://www.w3.org/2001/XMLSchema#int

Basic Ideas behind RDF

RDF uses Web identifiers (URIs) to identify resources

RDF builds relationships between resources

RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms



Ref: Linked Data Engineering , Prof. Dr. Harald Sack, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure & Karlsruhe Institute of Technology

http://dbpedia.org/page/Pluto

About: Pluto

An Entity of Type : planet, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Pluto (minor-planet designation: 134340 Pluto) is a dwarf planet in the Kuiper belt, a ring of bodies
the first Kuiper belt object to be discovered. It is the largest and second-most-massive known dwarf
System and the ninth-largest and tenth-most-massive known object directly orbiting the Sun. It is th
Neptunian object by volume but is less massive than Eris, a dwarf planet in the scattered disc. Like
objects, Pluto is primarily made of ice and rock and is relatively small—about one-sixth the mass of
third its volume. It has a moderately eccentric and inclined orbit during which it ranges from 30 to 4
AU (4.4–7.4 billion km) from the Sun. This means that Pluto perio

Property	Value
dbo:Planet/apoapsis	▪ 3.162498986598E11
dbo:Planet/averageSpeed	▪ 4.67
dbo:Planet/maximumTemperature	▪ 55.0
dbo:Planet/meanTemperature	▪ 44.0
dbo:Planet/minimumTemperature	▪ 33.0
dbo:Planet/periapsis	▪ 7.479893535E8

Dbpedia

- **DBpedia** is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web.

- Try out !
- http://dbpedia.org/resource/Pluto
- http://dbpedia.org/page/Pluto
- http://dbpedia.org/data/Pluto

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:prov="http://www.w3.org/ns/prov#"
    xmlns:ns4="http://dbpedia.org/ontology/Planet/"
    xmlns:dbo="http://dbpedia.org/ontology/"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:dbp="http://dbpedia.org/property/"
    xmlns:dct="http://purl.org/dc/terms/" >
  <rdf:Description rdf:about="http://dbpedia.org/resource/Clyde_Tombaugh">
    <dbo:knownFor rdf:resource="http://dbpedia.org/resource/Pluto" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Planet_Pluto">
    <dbo:wikiPageRedirects rdf:resource="http://dbpedia.org/resource/Pluto" />
  </rdf:Description>……————
```

<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> “1930” .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer> <http://dbpedia.org/resource/Clyde_Tombaugh> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/CelestialBody> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/place> .
… … …
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthdate> “1906-02-04” .
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthplace> <http://dbpedia.org/resource/Streator,_Illinois> .
… … …
<http://dbpedia.org/resource/Streator,_Illinois> <http://dbpedia.org/ontology/motto> “Quiet Surprise in the Prairie” .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#lat> “41.120834”^^xsd:float .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#long> “-88.835281”^^xsd:float .

RDF Triples: Subject – Property- Object

<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> “1930” .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer>
<http://dbpedia.org/resource/Clyde_Tombaugh> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/CelestialBody> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/place> .
… … ...
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthdate> “1906-02-04” .
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthplace>
<http://dbpedia.org/resource/Streator,_Illinois> .
… … ...
<http://dbpedia.org/resource/Streator,_Illinois> <http://dbpedia.org/ontology/motto> “Quiet Surprise in the Prairie” .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#lat> “41.120834”^^xsd:float .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#long> “-88.835281”^^xsd:float .

Individuals (Entities)

<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> “1930” .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer>
<http://dbpedia.org/resource/Clyde_Tombaugh> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/CelestialBody> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/place> .
… … ...
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthdate> “1906-02-04” .
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthplace>
<http://dbpedia.org/resource/Streator,_Illinois> .
… … ...
<http://dbpedia.org/resource/Streator,_Illinois> <http://dbpedia.org/ontology/motto> “Quiet Surprise in the Prairie” .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#lat> “41.120834”^^xsd:float .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#long> “-88.835281”^^xsd:float .

Classes

<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> “1930” .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer>
<http://dbpedia.org/resource/Clyde_Tombaugh> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/CelestialBody> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/place> .
… … ...
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthdate> “1906-02-04” .
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthplace>
<http://dbpedia.org/resource/Streator,_Illinois> .
… … ...
<http://dbpedia.org/resource/Streator,_Illinois> <http://dbpedia.org/ontology/motto> “Quiet Surprise in the Prairie” .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#lat>
“41.120834”^^xsd:float .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#long> “-88.835281”^^xsd:float .

Literals

<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> “1930” .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer>
<http://dbpedia.org/resource/Clyde_Tombaugh> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/CelestialBody> .

<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/place> .
… … ...
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthdate> “1906-02-04” .
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthplace>
<http://dbpedia.org/resource/Streator,_Illinois> .
… … ...
<http://dbpedia.org/resource/Streator,_Illinois> <http://dbpedia.org/ontology/motto> “Quiet Surprise in the Prairie” .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#lat> “41.120834”^^xsd:float .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#long> “-88.835281”^^xsd:float .

Vocabularies / Ontologies

RDF Serializations: Most popular formats

- Various serialization formats for different purposes are:

  N-Triples – a text format focusing on simple parsing
- **Turtle** – a text format focusing on human readability

- **Notation 3 (N3)** – a text format with advanced features beyond RDF
- **RDF/XML** – the official XML serialization of RDF
- **JSON-LD** – the official JSON serialization of RDF (supersedes earlier alternative approaches, e.g. RDF/JSON)
- **RDFa** – a mechanism for embedding RDFa in (X)HTML

N-Triples

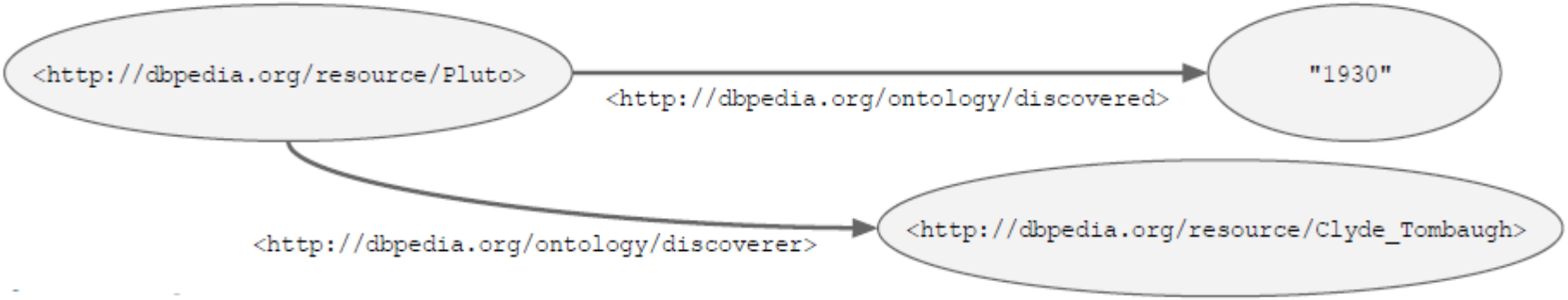
- N-Triples is a line-based, plain text format (http://www.w3.org/TR/n-triples/)

- N-Triples is a **subset of Turtle and Notation 3**
  - Abbreviations and grouping not allowed
  - Limited to ASCII character set

- **N-Triples Notation**
- ○ **URIs/IRIs** in angle brackets
- ○ **Literals** in quotation marks
- ○ Triple ends with a **period**

```
<http://www.w3.org/2001/sw/RDFCore/ntriples/>
       <http://purl.org/dc/elements/1.1/creator> "Dave Beckett" .

<http://www.w3.org/2001/sw/RDFCore/ntriples/>
       <http://purl.org/dc/elements/1.1/creator> "Art Barstow" .

<http://www.w3.org/2001/sw/RDFCore/ntriples/>
       <http://purl.org/dc/elements/1.1/publisher> <http://www.w3.org/> .
```

N-Triples



<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> "1930" .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer>
<http://dbpedia.org/resource/Clyde_Tombaugh> .

Turtle Syntax

- Turtle – **Terse RDF Triple Language** (subset of N3)
- URIs in angle brackets: *<http://dbpedia.org/resource/Berlin>*
- Literals in quotes:
  - *"Berlin"@de*
  - *"51.333332"^^xsd:float*
- A triple is terminated by a dot.
- White spaces and line breaks are ignored outside of identifiers
- Status: W3C Recommendation 25 February 2014 , http://www.w3.org/TR/turtle/

```
<http://dbpedia.org/resource/Aachen>
        <http://www.w3.org/2000/01/rdf-schema#label>
        "Aachen"@de .
```

Turtle Syntax

In Turtle one can use abbreviations

Syntax: @prefix abbr ':' <URI> .
E.g. @prefix dbr: <http://dbpedia.org/resource/> .

One can transform

<http://dbpedia.org/resource/Aachen> <http://www.w3.org/2000/01/rdf-schema#label> “Aachen"@de .

Into

@prefix dbr: <http://dbpedia.org/resource/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema> .
dbr:Aachen rdfs:label “Aachen"@de .

Turtle Syntax

Triples with the same subject can be grouped together

- **semicolon** indicates that subsequent triples have the same subject **(predicate list)**

```
@prefix rdf:
@prefix geo:
dbr:Leipzig dbp:hasMayor dbr:Burkhard_Jung ;
        rdfs:label   "Leipzig"@de ;
        geo:lat      "51.333332"^^xsd:float ;
        geo:long     "12.383333"^^xsd:float .
```

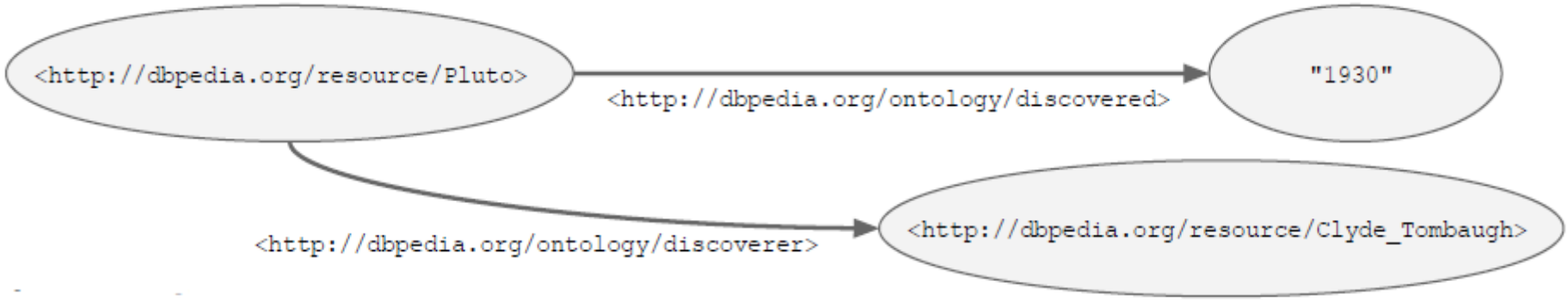
Turtle Syntax

triples with the same subject and predicate can be grouped together:

- **comma** indicates that subsequent triples have same subject and property **(object list)**

```
@prefix dbr:  .
@prefix dbp:  .
dbr:Leipzig dbp:locatedIn dbr:Saxony, dbr:Germany;
        dbp:hasMayor  dbr:Burkhard_Jung .
```

Turtle Notation



@prefix dbo: <http://dbpedia.org/ontology/> .
@base <http://dbpedia.org/resource/> .
<Pluto> dbo:discovered "1930" .
<Pluto> dbo:discoverer <Clyde_Tombaugh> .

RDF syntax

- Starting with <rdf:RDF> and end with </rdf:RDF>
- <rdf:Description> is the main element to define the subject, predicate and object of the statement
- RDF Namespace
  - http://www.w3.org/1999/02/22-rdf-syntax-ns#,
- File format: .rdf

RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

RDF-XML Example

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

<rdf:Description
  rdf:about="http://www.rechshop.fake/cd/Empire Burlesque">
    <cd:artist>Bob Dylan</cd:artist>
    <cd:country>USA</cd:country>
    <cd:company>Columbia</cd:company>
    <cd:price>10.90</cd:price>
    <cd:year>1985</cd:year>
</rdf:Description>
<rdf:Description
  rdf:about="http://www.rechshop.fake/cd/Hide your heart">
    <cd:artist>Bonnie Tyler</cd:artist>
    <cd:country>UK</cd:country>
    <cd:company>CBS Records</cd:company>
    <cd:price>9.90</cd:price>
    <cd:year>1988</cd:year>
</rdf:Description>
<!-- more cds -->
</rdf:RDF>
```

RDF main elements

- <rdf:RDF>: the root element
- <rdf:Description>: defining a resource

RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

<rdf:RDF>

- It is the root element of an RDF document
- It declares the XML document to be an RDF document
- It contains a reference to the RDF namespace

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

  .  .  .

</rdf:RDF>
```

<rdf:Description>

- It defines a resource using “about” attribute
- It contains elements that describing the resource (property, property values)

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

<rdf:Description
  rdf:about="http://www.rechshop.fake/cd/Empire Burlesque">
    <cd:artist>Bob Dylan</cd:artist>
    <cd:country>USA</cd:country>
    <cd:company>Columbia</cd:company>
    <cd:price>10.90</cd:price>
    <cd:year>1985</cd:year>
</rdf:Description>

</rdf:RDF>
```

rdf:about and rdf:ID

- Both are attribute for rdf:Description to represent the subject of the statement.
  - If the subject is complete URI, then use rdf:about
  - If the subject is fragment, then use rdf:ID
  - If the subject is a blank node, then use rdf:nodeID

```
<rdf:Description
  rdf:about="http://www.rechshop.fake/cd/Empire Burlesque">
</rdf:Description>

<rdf:Description
  rdf:ID="Empire Burlesque">
</rdf:Description>

<rdf:Description rdf:nodeID="abc">
</rdf:Description>
```


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

rdf:about and rdf:ID

- You can use a relative URI in rdf:about and resolve it based on the base URI.

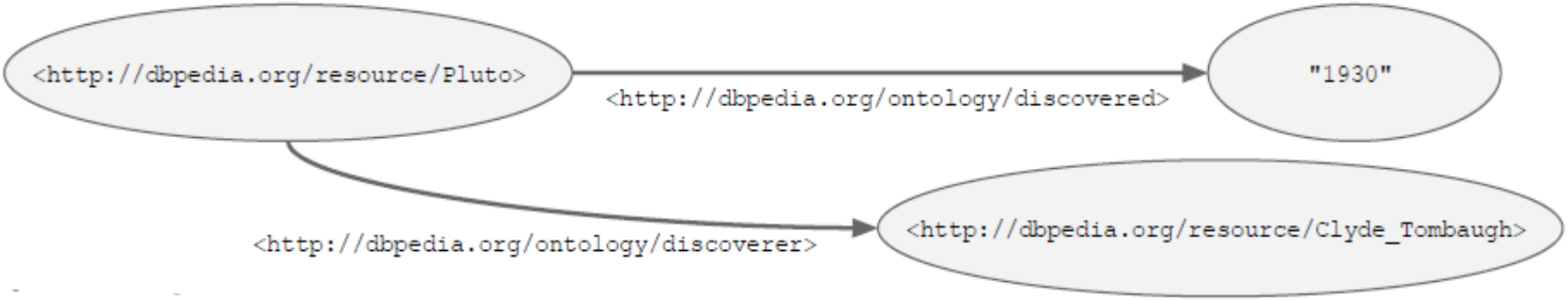
```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xml:base="http://spam.com/eggs/" >

<rdf:Description rdf:about="listing.rdf#local-record">
<dc:title>Local Record</dc:title>
</rdf:Description>
</rdf:RDF>
```

The whole URI for local-record is:

http://spam.com/eggs/listing.rdf#local-record

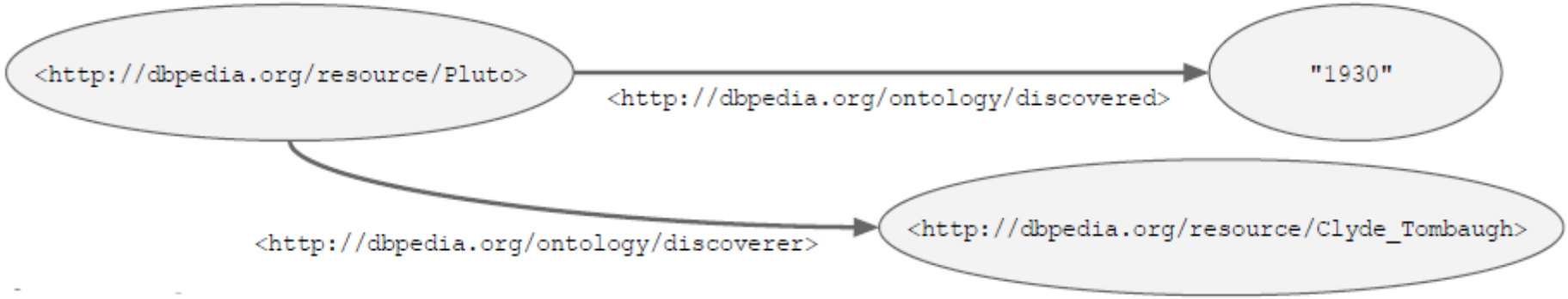
RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

RDF/XML Notation




```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                xmlns:ns0="http://dbpedia.org/ontology/">
<rdf:Description rdf:about="http://dbpedia.org/resource/Pluto">
    <ns0:discovered>1930</ns0:discovered>
    <ns0:discoverer rdf:resource="http://dbpedia.org/resource/Clyde_Tombaugh"/>
  </rdf:Description>
</rdf:RDF>
```

JSON-LD Notation (RDF 1.1)




```
{ "@id" : "http://dbpedia.org/resource/Pluto" ,
      "http://dbpedia.org/ontology/discovered" :
            { "@value" : "1930" }
   ,
      "http://dbpedia.org/ontology/discoverer" :
            { "@id" : "http://dbpedia.org/resource/Clyde_Tombaugh"}
}
```

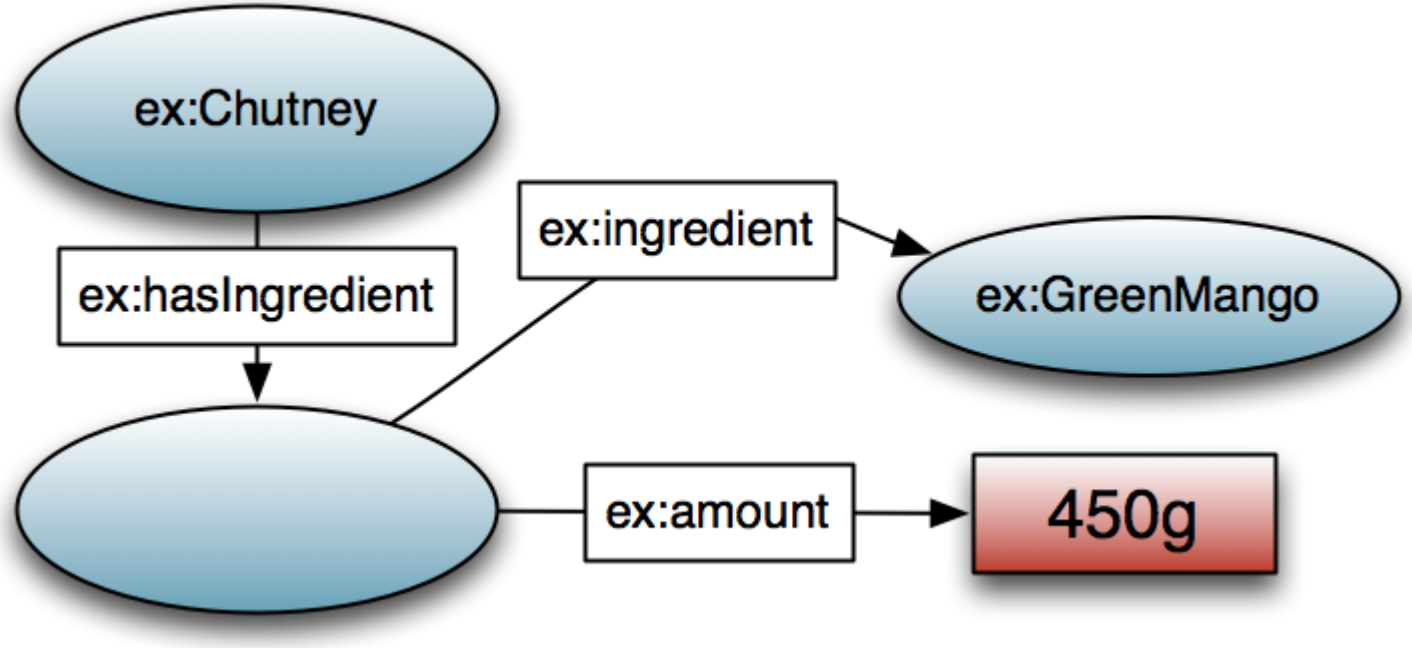
Blank Nodes

- denote existence of an individual with specific attributes, but without providing an identification or reference

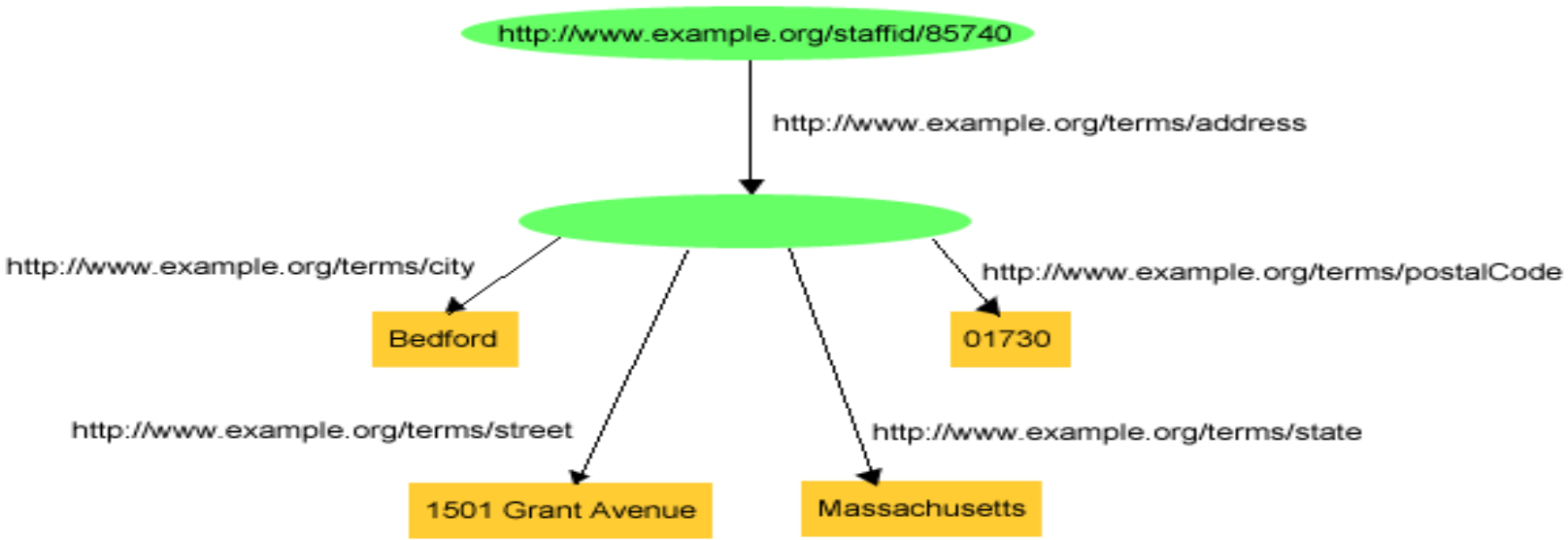
- A blank node has no node identifier (has no name), but
  - Convention provides a way to use a blank node identifier to distinguish blank nodes from other nodes.
  - When merging different RDF graphs, different blank nodes need to be distinctly identified.

Use of Blank Nodes

- . "For the preparation of mango chutney you need 450g of green mango , a teaspoon of cayenne pepper ..."



Blank Node

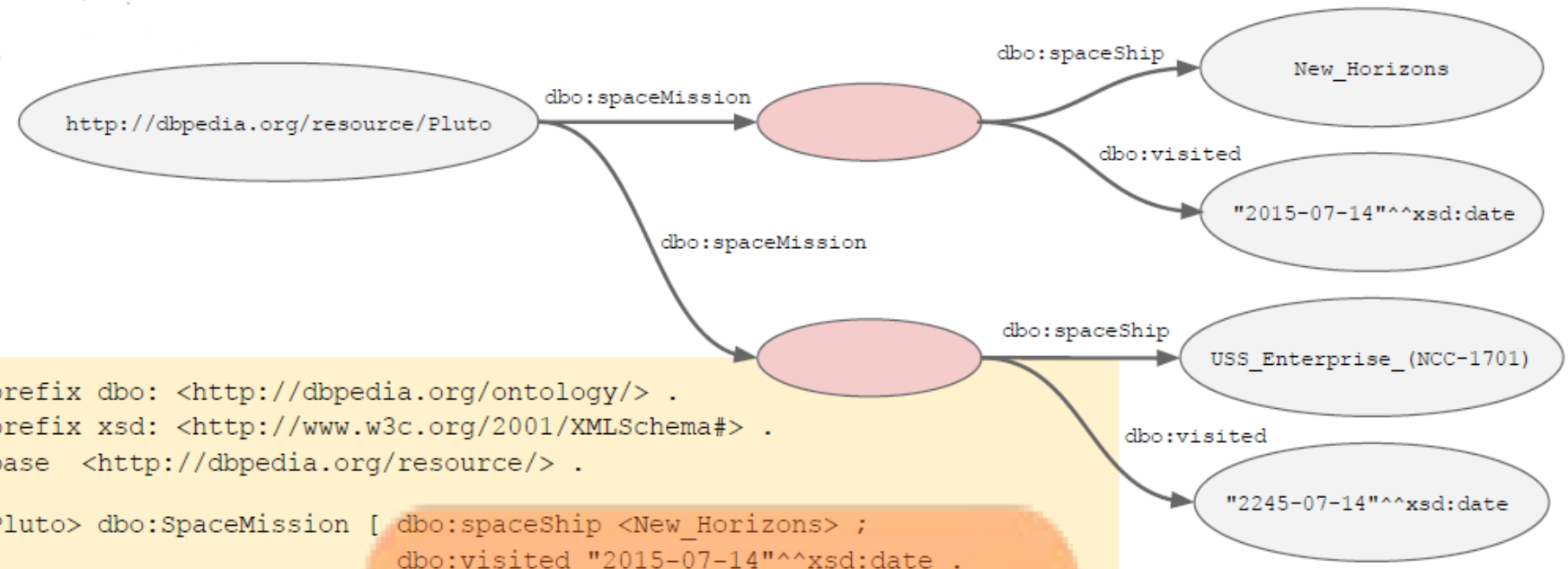


If we gives _:johnaddress as the identifier for the blank node

```
exstaff:85740 exterms:address _:johnaddress .
_:johnaddress exterms:street "1501 Grant Avenue" .
_:johnaddress exterms:city "Bedford" .
_:johnaddress exterms:state "Massachusetts" .
_:johnaddress exterms:postalCode "01730".
```


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Blank Nodes: Turtle - Terse RDF Triple Language




```
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix xsd: <http://www.w3c.org/2001/XMLSchema#> .
@base  <http://dbpedia.org/resource/> .

<Pluto> dbo:SpaceMission [ dbo:spaceShip <New_Horizons> ;
                           dbo:visited "2015-07-14"^^xsd:date .
                         ],
                         [ dbo:spaceShip <USS_Enterprise_(NCC-1701)> ;
                           dbo:visited "2245-07-14"^^xsd:date .
                         ].
```

nested anonymous blank nodes

Ref: Linked Data Engineering , Prof. Dr. Harald Sack, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure & Karlsruhe Institute of Technology

Lists

- General data structures for enumerating arbitrarily many resources
- Distinction between
  - **Container:** adding new elements possible

    ordered and unordered container types
  - **Collections:** ordered list; adding new elements impossible

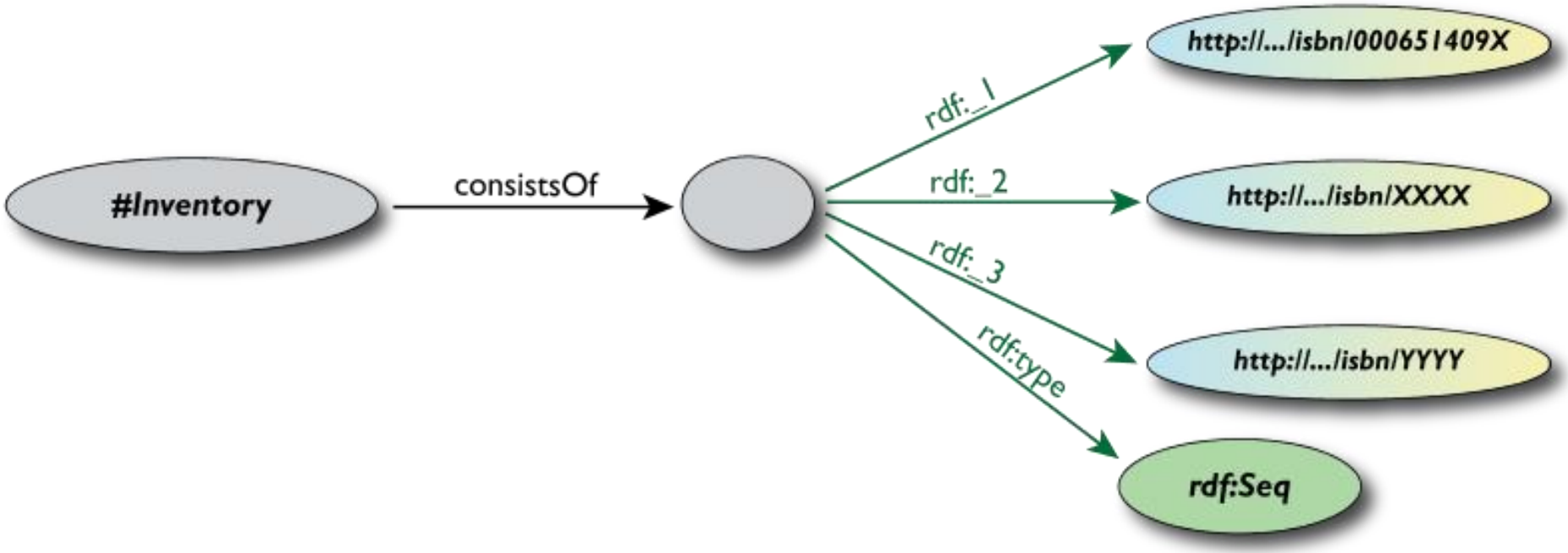
Types of Containers

- The list root node is assigned one of the following

rdf:types:

- **rdf:Seq**
  - Interpretation as ordered list, sequence
  - <Seq>: a list of members with order
- **rdf:Bag**
  - Interpretation as unordered set
  - Order coded in RDF not relevant
  - <Bag>: a list of members without order
- **rdf:Alt**
  - Set of alternatives
  - Usually only one list element is relevant
  - <Alt>: a list of members that only one can be selected

Example


#Inventory
consistsOf
rdf:_1
rdf:_2
rdf:_3
rdf:type
http://.../isbn/000651409X
http://.../isbn/XXXX
http://.../isbn/YYYY
rdf:Seq

<rdf:Bag>

- It is used to describe a list of values without order
- It can contain duplicate values

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cd="http://www.rechshop.fake/cd#">
<rdf:Description rdf:about="http://www.rechshop.fake/cd/Beatles">
  <cd:artist>
    <rdf:Bag>
      <rdf:li>John</rdf:li>
      <rdf:li>Paul</rdf:li>
      <rdf:li>George</rdf:li>
      <rdf:li>Ringo</rdf:li>
    </rdf:Bag>
  </cd:artist>
</rdf:Description>
</rdf:RDF>
```


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

<rdf:Seq>

- It is used to describe a list of values with order
- It can contain duplicate values

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cd="http://www.rechshop.fake/cd#">
<rdf:Description rdf:about="http://www.rechshop.fake/cd/Beatles">
  <cd:artist>
    <rdf:Seq>
      <rdf:li>John</rdf:li>
      <rdf:li>Paul</rdf:li>
      <rdf:li>George</rdf:li>
      <rdf:li>Ringo</rdf:li>
    </rdf:Seq>
  </cd:artist>
</rdf:Description>
</rdf:RDF>
```


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

rdf:Seq

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org/test#> .
    ex:SolarSystem ex:planets [
   a rdf:Seq ;
   rdf:_1 ex:Mercury ;
   rdf:_2 ex:Venus ;
   rdf:_3 ex:Earth ;
   rdf:_4 ex:Mars ;
   rdf:_5 ex:Jupiter ;
   rdf:_6 ex:Saturn
 ] .
```

<rdf:Alt>

- It is used to describe a list of alternative values that the user can select only one of it

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

<rdf:Description rdf:about="http://www.recshop.fake/cd/Beatles">
  <cd:format>
    <rdf:Alt>
      <rdf:li>CD</rdf:li>
      <rdf:li>Record</rdf:li>
      <rdf:li>Tape</rdf:li>
    </rdf:Alt>
  </cd:format>
</rdf:Description>

</rdf:RDF>
```


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

rdf:parseType="Collection"

- It is used to describe group that contains ONLY the specified members
- It is described as the attribute rdf:parseType="Collection"

- rdf:parseType="Collection": enumerates the specified members (the group only contains the specified members listed in the collection)

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cd="http://www.rechshop.fake/cd#">
<rdf:Description rdf:about="http://www.rechshop.fake/cd/Beatles">
 <cd:artist rdf:parseType="Collection">
  <rdf:Description rdf:about="http://recshop.fake/cd/Beatles/George"/>
  <rdf:Description rdf:about="http://recshop.fake/cd/Beatles/John"/>
  <rdf:Description rdf:about="http://recshop.fake/cd/Beatles/Paul"/>
  <rdf:Description rdf:about="http://recshop.fake/cd/Beatles/Ringo"/>
 </cd:artist>
</rdf:Description>
</rdf:RDF>
```


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

Turtle : Collection

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org/test#> .
  ex:SolarSystem ex:planets (
  ex:Mercury ex:Venus ex:Earth ex:Mars ex:Jupiter ex:Saturn
 ) .
```

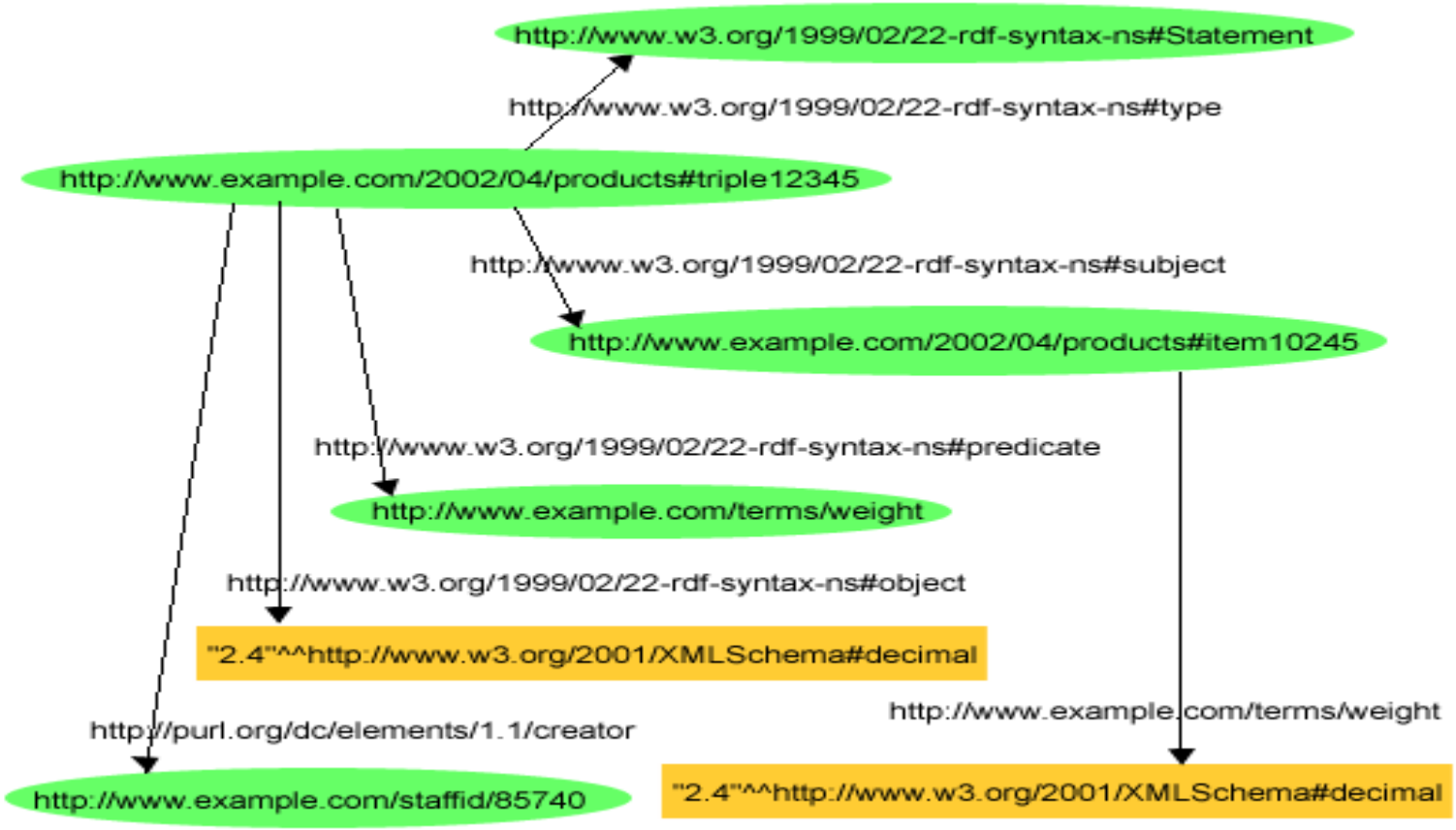
RDF Reification

- RDF provides a built-in vocabulary for describing RDF statements.
- A description of a statement using this vocabulary is called a reification of the statement
  - rdf:Statement, rdf:subject, rdf:predicate, rdf:object

```
<rdf:Statement rdf:about="#triple12345">
  <rdf:subject
        rdf:resource="http://www.example.com/2002/04/products#item10
        245"/>
  <rdf:predicate
        rdf:resource="http://www.example.com/terms/weight"/>
  <rdf:object rdf:datatype="&xsd;decimal">2.4</rdf:object>

  <dc:creator rdf:resource="http://www.example.com/staffid/85740"/>

</rdf:Statement>
```


RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

RDF Reification




```
exproducts:triple12345 rdf:type rdf:Statement .
exproducts:triple12345 rdf:subject exproducts:item10245 .
exproducts:triple12345 rdf:predicate exterms:weight .
exproducts:triple12345 rdf:object "2.4"^^xsd:decimal .
exproducts:triple12345 dc:creator exstaff:85740 .
```

RDF Reification

- Sherlock Holmes supposes that the gardener has killed the butler

- **Part 1**: the gardener has killed the butler
- ex:Gardener ex:hasKilled ex:butler .
- **Part 2**: Sherlock Holmes supposes...
- dbpedia:Sherlock_Holmes ex:supposes

@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex: <http://example.org/Crimestories#> .
dbpedia:SherlockHolmes **ex:supposes** ex:StatementOfSherlock .
ex:Gardener ;
    rdf:predicate ex:hasKilled ;
    rdf:object ex:Butler .

Biomedical RDF Data

- Endpoints:

- Uniport
- http://sparql.uniprot.org/

- http://www.uniprot.org/

- Chembl
- https://www.ebi.ac.uk/chembl/

- https://www.ebi.ac.uk/rdf/services/chembl/sparql

- Atlas RDF

- https://www.ebi.ac.uk/rdf/services/atlas/

- https://www.ebi.ac.uk/rdf/services/atlas/sparql