2nd Annual CHANGES Workshop
Data Driven Science: "Data Management, Analytics and Visualization"

Publishing and Linking Scientific Data

—*Experience on Applying Linked Data to Scientific Database Project*

SHEN Zhihong
Computer Network Information Center, CAS
2013/9/10

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Outline

- Background
- Publishing scientific data
- Linking scientific data
- Summary
- Next work

Background - Scientific DataBase project

- Scientific DataBase (SDB) project
- A long-term mission funded by CAS(Chinese Academy of Sciences) which started in 1986
  - data from research, for research
- Collecting multi-discipline research data and promoting data sharing (2006-2010, during the period of the eleventh Five-Year-Plan of CAS)
  - about **61** CAS institutes involved
  - Over **200TB** data available for open access and download

http://www.csdb.cn

Background - Scientific Databases

- SDB consists of 51 databases, including …

- 8 Resource databases
  - Geo-Science
  - Biodiversity
  - Chemistry
  - Astronomy
  - Space Science
  - Micro biology and virus
  - Material science
  - Environment
- 37 institution databases

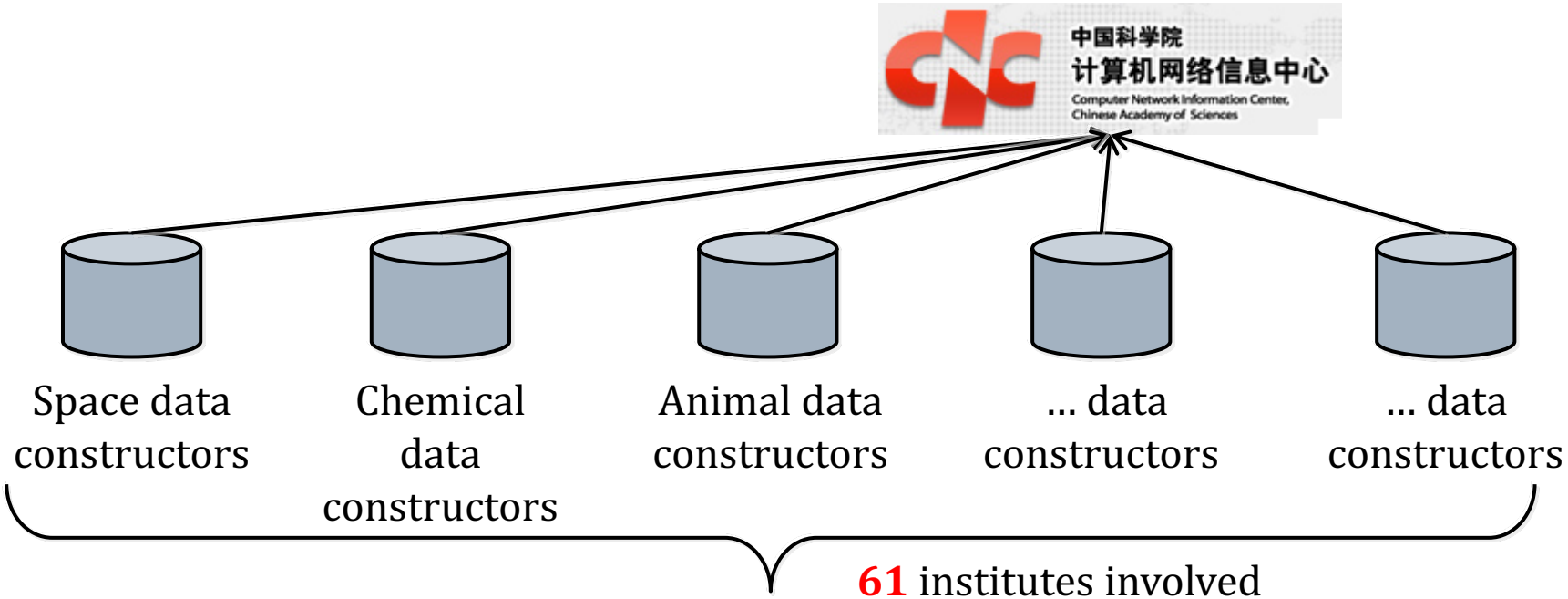
- 2 Reference databases
  - China Species
  - chemical compound
- 4 Application-Oriented databases
  - High Energy (ITER)
  - Western Environment Research
  - Ecology research
  - Qinghai Lake Research






Background

- Role of CNIC in SDB project
  1. plays the role of the organizer of the project
  2. offers the cyberinfrastructure for the project
     - especially data storage environment
  3. develops generic tools for building database systems, and develops SDB portal



Background

- Big question to CNIC
  - How to use all data involved and provide unified services for users? Not just a mess
- One possible idea is …
  - **Put all data together into CNIC**?
  - This idea is very effective, but always means painful.
    - Data ownership: data is owned by different institutes, CNIC has no right to own these data and publish them
    - Inborn heterogeneity: different data, different format, different software environments …
    - Keep data updated: a difficult task

Background

- **Data linking** makes sense
  - Data ownership: data is published in strict access control by data owners, we just link them!
  - Heterogeneity: data keeps original state, Data linking means no additional works for data owners
  - Keep data updated: data is updated by owners
- So, the best idea for data integration is
  - **Data Linking**
  - Not **moving** all data together, but **linking** them together

Background

- How we do Data Linking
  1. Data owners **publishes** scientific data and scientific metadata in a good format
  2. CNIC Collects all allowed data descriptions and provides **searching** interfaces for users
  3. CNIC generates and discovers **links** among data and shows related links according to user query

Background

- key problems to be solved
  1. **Publishing**: A good format? What is a good format? Some data has little information, how to publish them?
  2. **Link generation**: really need automatic methods and runnable frameworks, especially for a large volume of data

Background

Publishing scientific data

- RDF and Linked Data as good formats
- Publishing data records
- Publishing data files

Linking scientific data

Summary

Next Work

Publishing scientific data

- What is a good format for scientific data?
  1. Be able to represent structured data and semi-structured data
     - Relation records, MathML, CML, SMILES, ...
  2. It is impossible to invent a new uniform format for all unstructured scientific data, then how about to be a good format for metadata?
     - Metadata of NetCDF, HDF, …

Data types	examples	requirements
Structured data	Relational data	Be able to represent
Semi-structured data	MathML, CML, SMILES	Be able to represent
Unstructured data	NetCDF, HDF	Be able to represent its metadata

Publishing scientific data

- What is a good format for scientific data?
  3. able to describe the links between data
  4. able to identify each data on the Web
  5. machine-readable and understandable
     - May be consumed by programs
  6. flexible schema, easy to extend, adding a new property is very simple

Publishing scientific data

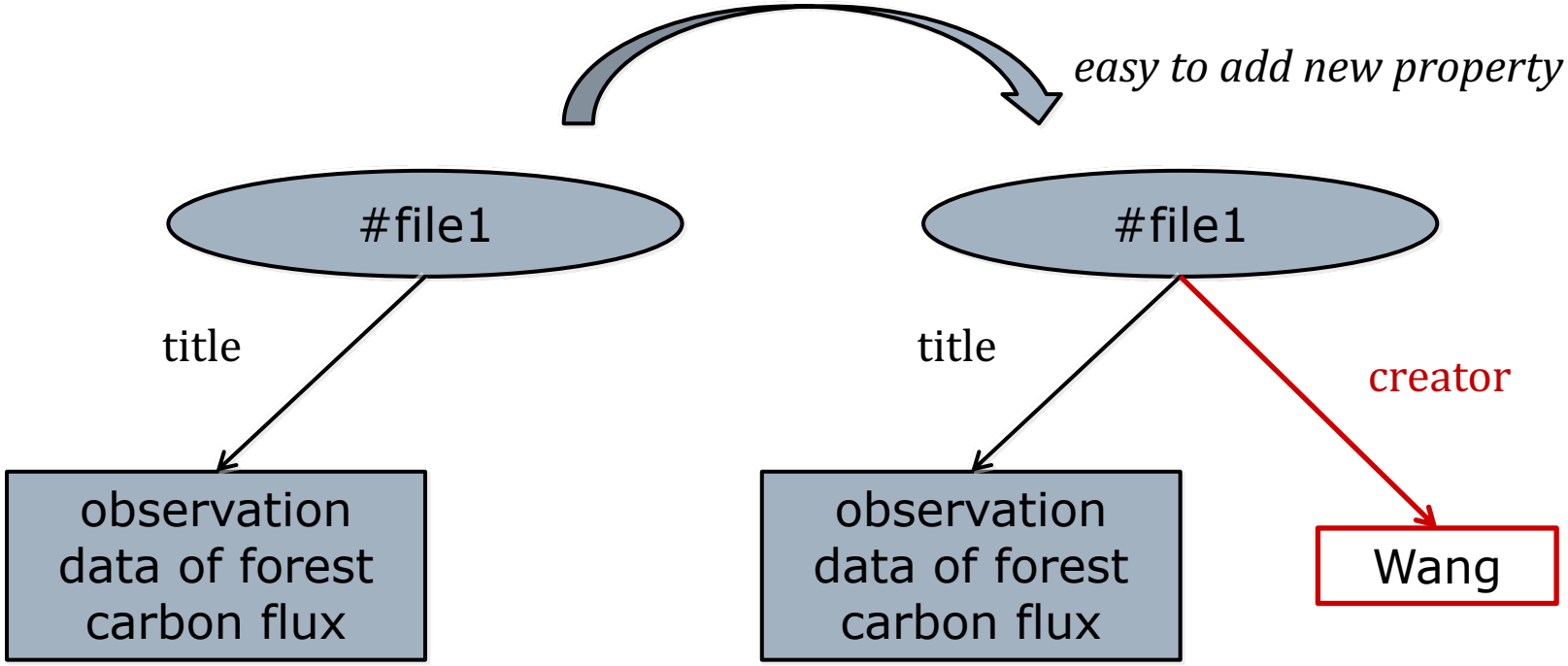
- We chose RDF as data representation format
  - RDF: **Resource Description Framework**
  - RDF describes resources in terms of simple properties and property values
  - it is based upon the idea of *making statements about resources* (in particular web resources) in the form of subject-predicate-object expressions
  - e.g

<#dataSet1> <title> "observation data of forest carbon flux"

subject predicate object

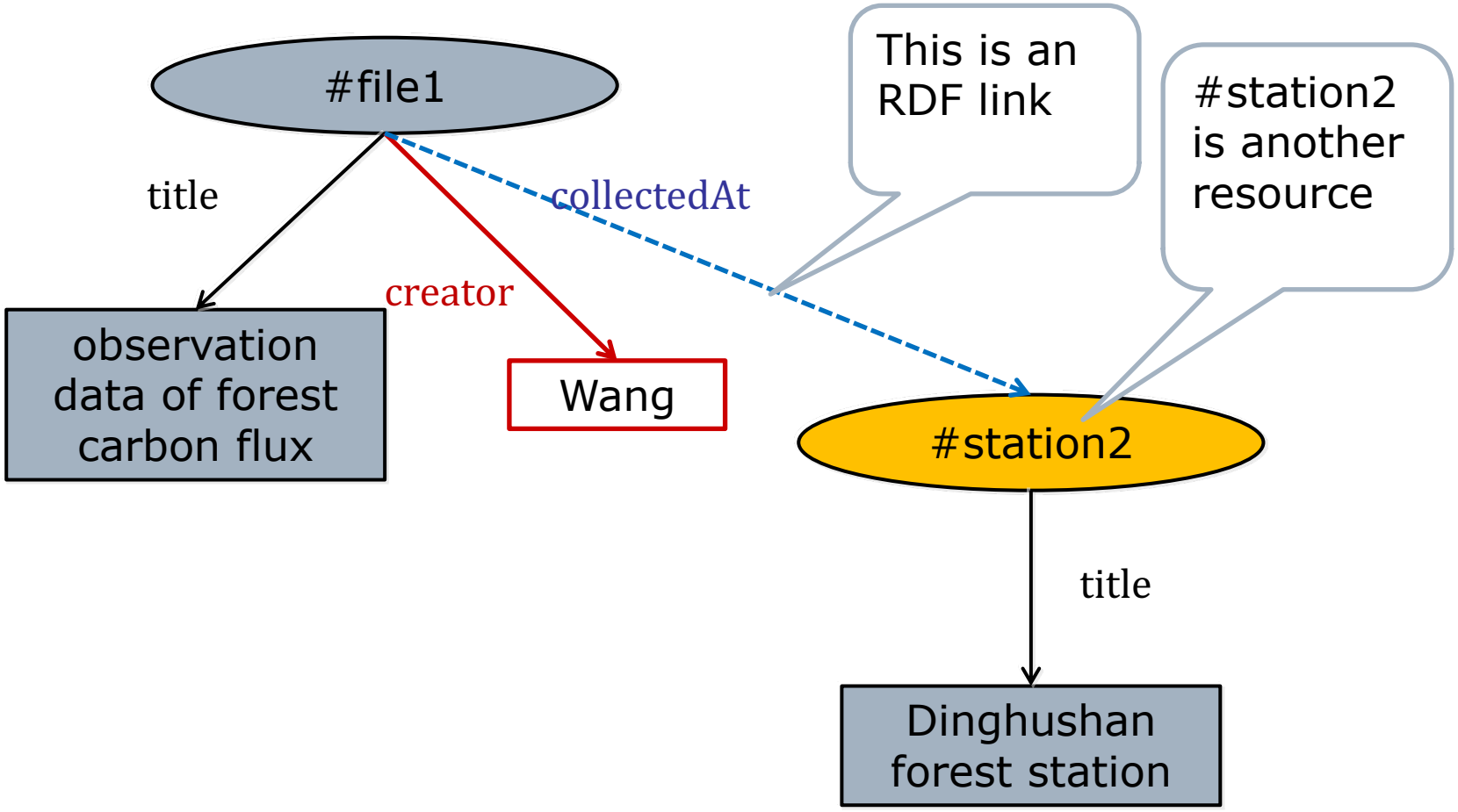
Publishing scientific data

- RDF statements about a resource are often represented as a graph
  - nodes: representing the resources, and their properties values.
  - arcs: representing properties of resources



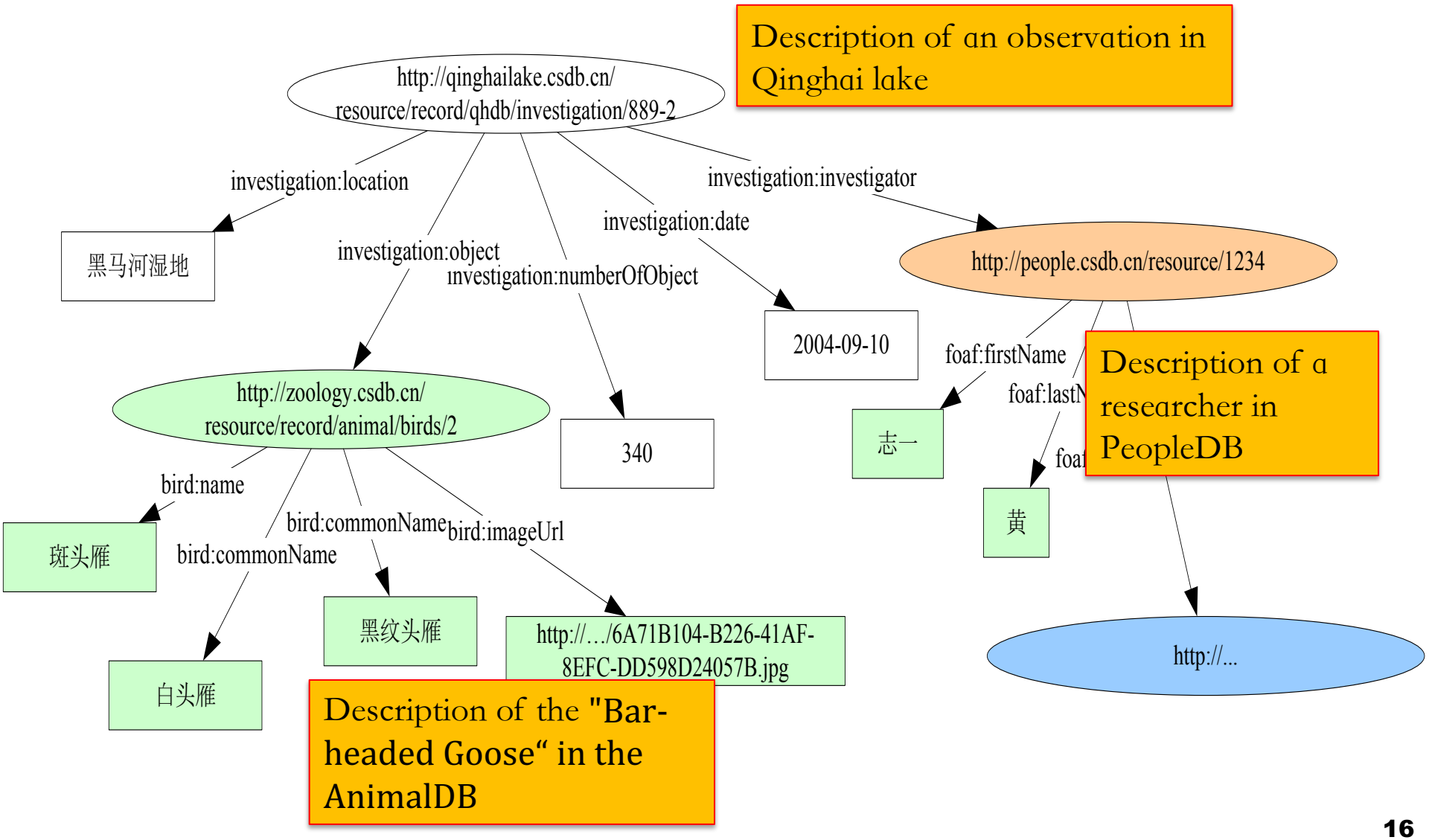
Publishing scientific data

- Most important, RDF data model enables people to set RDF links between data from different sources.



Publishing scientific data

- More complex RDF data example



Publishing scientific data

- After choosing the data format, we chose Linked Data as the interoperation mechanism of data sources
  - Linked Data: Tim Berners-Lee coined the term Linked Data in 2006[1].
  - Connect Distributed Data across the Web
  - using the Web to create typed links between data from different sources

1. T. Berners-Lee, "Design issues: Linked data," Online at http://www. w3. org/DesignIssues/LinkedData. html, 2006

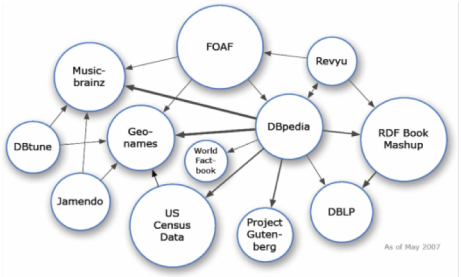
Publishing scientific data

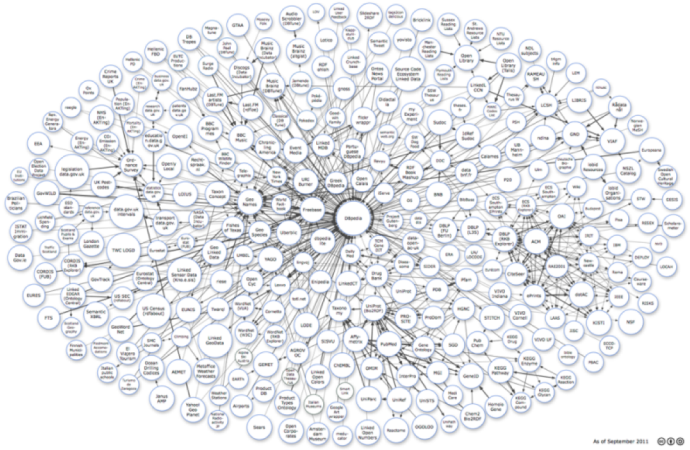
- 4 basic Linked Data Principles
  1. Use URIs as names for things
  2. Use HTTP URIs so that people can look up (dereference) those names.
  3. When someone looks up a URI, provide useful information.
  4. Include links to other URIs so that they can discover more things.

Publishing scientific data

LINKINGOPENDATA
W3C SWEO Community Project

- Linking Open Data
  - Based on the concept of linked data, W3C initiated the Linking Open Data movement.
  - It has driven many data sets published as Linked Data.
  - By September 2011, LOD had covered about 295 datasets with 25 billion RDF triples and about 395 million RDF links.





1. "Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

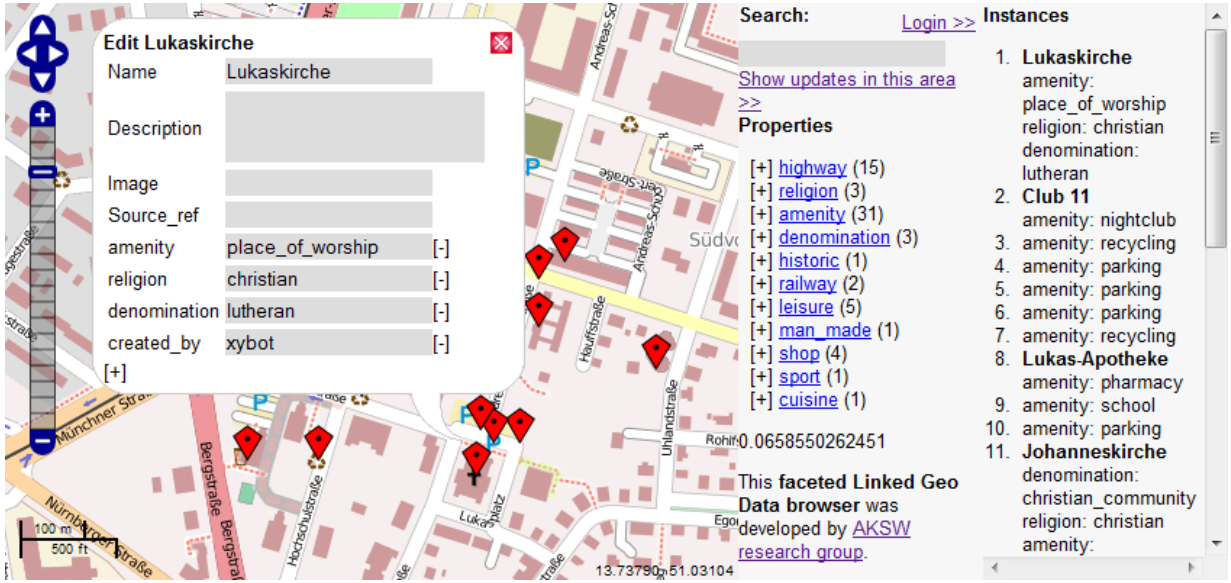
Publishing scientific data

- Scientific Data samples – Linked Life Data
  - Searches and explores over RDF statements from various sources including UniProt, PubMed, EntrezGene and 20 more…
  - Performs complex SPARQL(RDF query) queries and retrieves more than one billion RDF resources.

1. Momtchev V, Peychev D, Primov T, et al. Expanding the pathway and interaction knowledge in linked life data[C]. In Proceedings of International Semantic Web Challenge, 2009.

Publishing scientific data

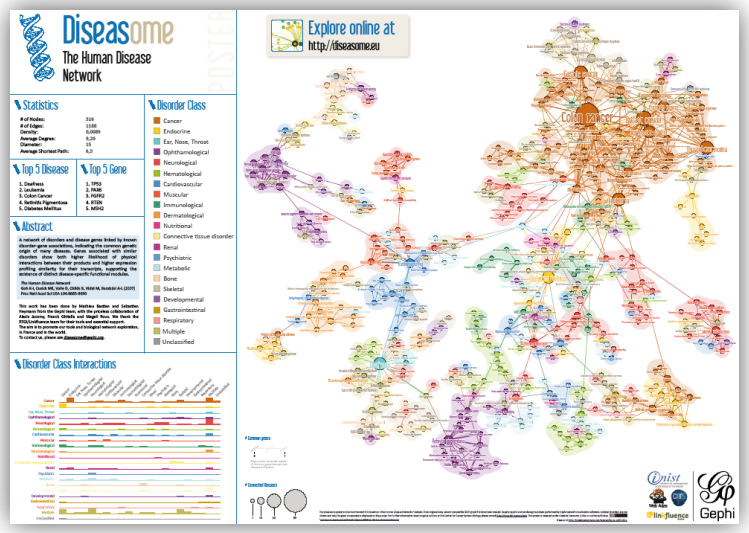
- Scientific Data samples – LinkedGeoData
  - uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles.



1. Auer, Sören, Jens Lehmann, and Sebastian Hellmann. "Linkedgeodata: Adding a spatial dimension to the web of data." The Semantic Web-ISWC 2009. Springer Berlin Heidelberg, 2009. 731-746.

Publishing scientific data

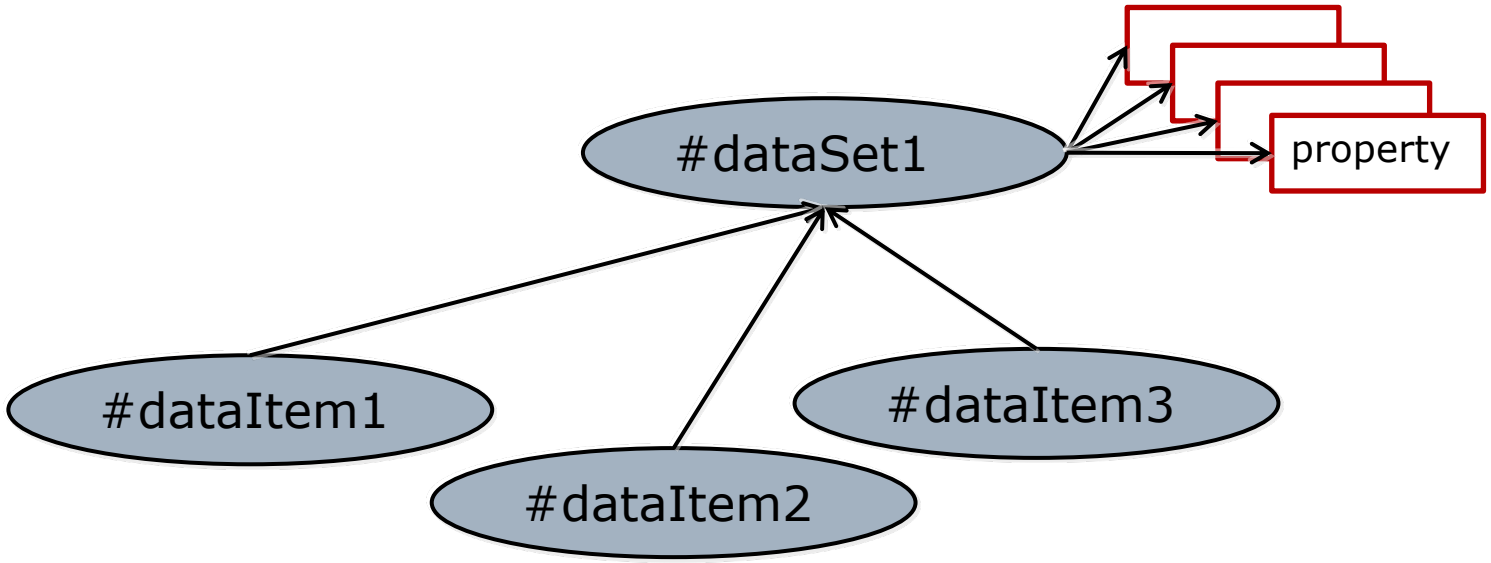
- Diseasome
  - publishes Linked Data of 4,300 disorders and **disease genes** linked by known disorder-gene associations for exploring all known phenotype and disease gene associations, indicating the common genetic origin of many diseases.
- Linked Sensor Data
  - the first open datasets for sensors and sensor observations, created at Knoesis Center, and converted from weather data at Mesowest.
  - Contains descriptions of 20 thousand weather stations and 160 million observations.
- GeoSpecies Knowledge Base
  - Publishing information on Biological Orders, Families, Species as well as species occurrence records and related data, links to geonames, bio2rdf, dbpedia, freebase, umbel.



1. Diseasome | Map: explore the human disease network. Dataset, interactive map and printable poster of gene-disease relationships. http://diseasome.eu/map.html
2. http://wiki.knoesis.org/index.php/SSW_Datasets

Publishing scientific data

- We define data in 2 levels
  1. Data item
  2. Data set: a set of data items
  - Some data items have little information, but the information of data set can help users to find them



Publishing scientific data

- Dataset metadata is described with mixed RDF vocabularies
  1. Dublin-Core
     - http://purl.org/dc/elements/1.1/
  2. DC-TERMS
     - DCMI Metadata Terms
     - http://purl.org/dc/terms#
  3. PRISM
     - Publishing Requirements for Industry Standard Metadata
     - http://prismstandard.org/namespaces/basic/2.0/

Publishing scientific data

An example of dataset metadata:

```
<csdb:Database rdf:about="http://semweb.csdb.cn/csdb/resource/database/12053084">
        <dc:title>蒋家沟降水观测资料</dc:title>
        <prism:publicationDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2010-12-
03T09:51:59</prism:publicationDate>
        <dc:subject>地球科学</dc:subject>
        <rdfs:label>蒋家沟降水观测资料</rdfs:label>
        <prism:url>http://nsl.imde.ac.cn;http://www.mountain.csdb.cn/page/showEntity.vpage?uri=mouhazards.cata
HazardObsdata</prism:url>
        <dc:source>中国科学院东川泥石流观测研究站</dc:source>
        <dcterms:accrualPeriodicity>year</dcterms:accrualPeriodicity>
        <dc:contributor>中国科学院东川泥石流观测研究站</dc:contributor>
        <dc:type>004.01</dc:type>
        <dcterms:rights>数据使用者必须与数据提供者签订数据使用共享协议，使用后必须注明数据来源。
</dcterms:rights>
        <csdb:dqInfo rdf:resource="http://semweb.csdb.cn/csdb/resource/dqinfo/31022770"/>
        <prism:creationDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2009-08-
08</prism:creationDate>
        <dcterms:rightsHolder rdf:resource="http://semweb.csdb.cn/csdb/resource/contact/30941721"/>
        <prism:keyword>泥石流;降水</prism:keyword>
        <dc:description>本数据集收录中国科学院东川泥石流观测研究站在云南蒋家沟观测到的降水资料。降水资料
包括5个观测站点的长期观测资料。</dc:description>
        <dc:creator>中国科学院水利部成都山地灾害与环境研究所</dc:creator>
        <csdb:sharePolicy>本数据保密期5年，解密后免费使用。</csdb:sharePolicy>
        <dc:coverage rdf:resource="http://semweb.csdb.cn/csdb/resource/coverage/17137914"/>
        <csdb:purpose>本数据集的降水观测资料可以与蒋家沟泥石流暴发资料配合使用，是研究泥石流形成和泥石
流预报的珍贵资料。</csdb:purpose>
</csdb:Database>
```

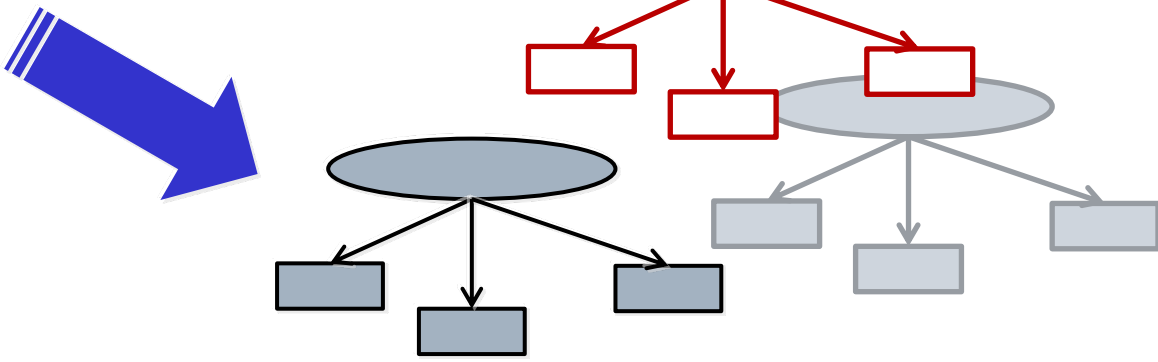
Publishing scientific data

- We distinguish data items into 2 classes
  - Data records:
    - records stored in databases
    - structured
  - Data files
    - file content and its metadata
    - File content is unstructured
    - Metadata is semi-structured

Publishing scientific data

- D2R mapper: Publishing data records
  - A table is mapped to a dataset
  - A record is mapped to an RDF resource
  - Each field is mapped to an RDF property
  - Field values are mapped to RDF property values

	r_regionkey integer	r_name character(25)	r_comment character varying(152)
1	0	AFRICA	lar deposits. blithely final packages ca
2	1	AMERICA	hs use ironic, even requests. s
3	2	ASIA	ges. thinly even pinto beans ca
4	3	EUROPE	ly final courts cajole furiously final e
5	4	MIDDLE EAST	uickly special accounts cajole carefully



Publishing scientific data

- F2R mapper: Publishing data files
  - File Content: binary-stream over HTTP (Non-RDF)
  - File Metadata
    - Physical information of a file
      - filename, size, creation time
    - Auto metadata extraction from scientific data files
      - FITS
      - HDF4
      - JPG
      - NetCDF
      - PowerPoint
      - Visio
      - Word

Publishing scientific data

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:j.0="http://www.semanticdesktop.org/ontologies/2007/05/10/nexif#"
    xmlns:j.1="http://www.csdb.cn/ns/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
  <rdf:Description rdf:nodeID="A0">
    <j.1:filepath>/test/图片1.JPG</j.1:filepath>
    <j.0:make>NIKON CORPORATION</j.0:make>
    <j.0:exposureBiasValue>5/3</j.0:exposureBiasValue>
    <rdf:type rdf:resource="http://www.semanticdesktop.org/ontologies/2007/05/10/nexif#Photo"/>
    <j.0:width>1680</j.0:width>
    <j.1:filename>图片1.JPG</j.1:filename>
    <rdfs:belongs>default</rdfs:belongs>
    <j.0:flashpixVersion>48 49 48 48</j.0:flashpixVersion>
    <rdfs:type>file</rdfs:type>
    <j.0:exposureProgram>1</j.0:exposureProgram>
    <j.1:filelastmodified>1317004200000</j.1:filelastmodified>
    <j.0:flash>0</j.0:flash>
    <j.0:height>2520</j.0:height>
    <j.1:filesize>2456762</j.1:filesize>
    <j.0:exposureMode>1</j.0:exposureMode>
    <j.1:filetype>file</j.1:filetype>
    <j.0:exposureTime>1/80</j.0:exposureTime>
  </rdf:Description>
</rdf:RDF>
```

extracted metadata of a JPEG file

extracted metadata of a FITS file

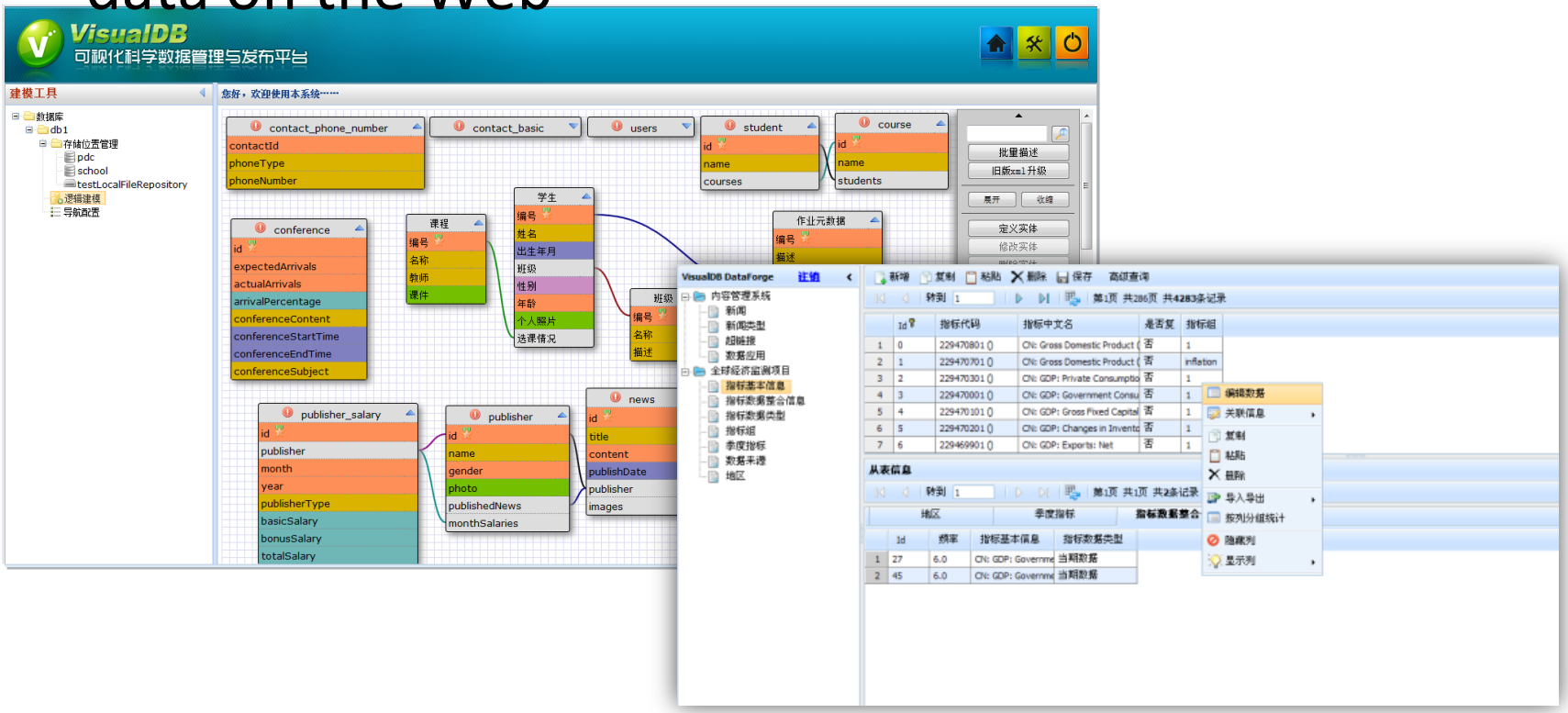
Problems | Javadoc | SVN 资源库 | Search | Console | Tasks | Debug | History | SVN 属性 | SVN

<terminated> FileMetaDataExtractionJobTest [JUnit] C:\Genuitec\Common\binary\com.sun.java.jdk.win32.x86_1.6.0.013\bin

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:j.0="http://www.csdb.cn/ns/"
    xmlns:j.1="http://www.semanticdesktop.org/ontologies/2007/03/22/nco#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
  <rdf:Description rdf:nodeID="A0">
    <j.1:CRVAL1>53.122751</j.1:CRVAL1>
    <j.1:PI-COI>C. Cesarsky</j.1:PI-COI>
    <j.0:filelastmodified>1317005340000</j.0:filelastmodified>
    <j.1:TEXPTIME>18000</j.1:TEXPTIME>
    <j.1:RA>53.161696</j.1:RA>
    <j.1:EXTEND>true</j.1:EXTEND>
    <j.0:filename>GOODS_ISAAC_03_H_V1.5.fits</j.0:filename>
    <j.1:MJDEND>53397.158255</j.1:MJDEND>
    <j.1:PHOTSYS>AB</j.1:PHOTSYS>
    <j.1:SEEINERR>0.04</j.1:SEEINERR>
    <j.1:NAXIS>2</j.1:NAXIS>
    <j.1:ORIGIN>GOODS</j.1:ORIGIN>
    <j.1:VORELDAT>30 Sep 2005</j.1:VORELDAT>
    <j.1:VERSION>1.5</j.1:VERSION>
    <j.1:GAIN>80999.999</j.1:GAIN>
    <j.1:CRPIX1>1396.5</j.1:CRPIX1>
    <j.1:CTYPE1>RA---TAN</j.1:CTYPE1>
    <j.1:OBSEND>2005-01-27T03:47:53</j.1:OBSEND>
    <j.1:ASTRERR>0.1</j.1:ASTRERR>
    <j.1:DEC>-27.706750</j.1:DEC>
    <j.1:ASTRCAT>GSC2</j.1:ASTRCAT>
    <j.1:FILTER>H</j.1:FILTER>
    <j.1:PHOTZPER>0.04</j.1:PHOTZPER>
    <j.1:PHOTZP>26.00</j.1:PHOTZP>
    <j.1:OBJECT>GOODS_03</j.1:OBJECT>
    <j.1:PROCSOFT>ESO/MVM</j.1:PROCSOFT>
    <rdfs:belongs>default</rdfs:belongs>
```

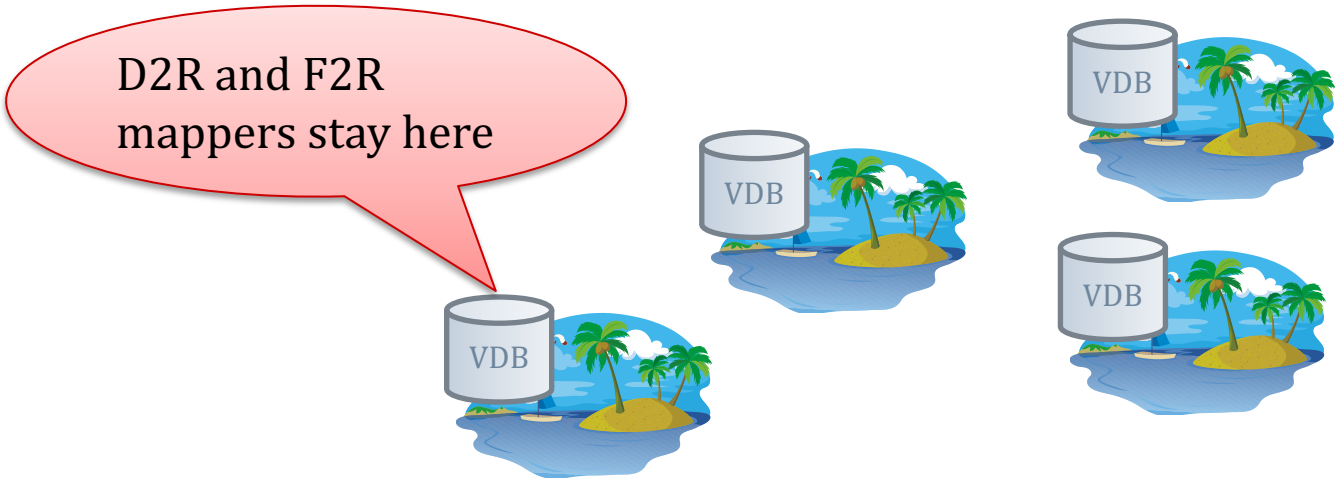
Publishing scientific data

- In SDB project, we embed D2R and F2R mappers in VisualDB
- VisualDB is a widely used tool developed by CNIC for helping data owners to manage and publish scientific data on the Web



Publishing scientific data

- VisualDB has been deployed in more than 30 institutes
  - 4.1 billion records
  - 26 million files
  - are published as Linked Data



Background

Publishing scientific data

Linking scientific data

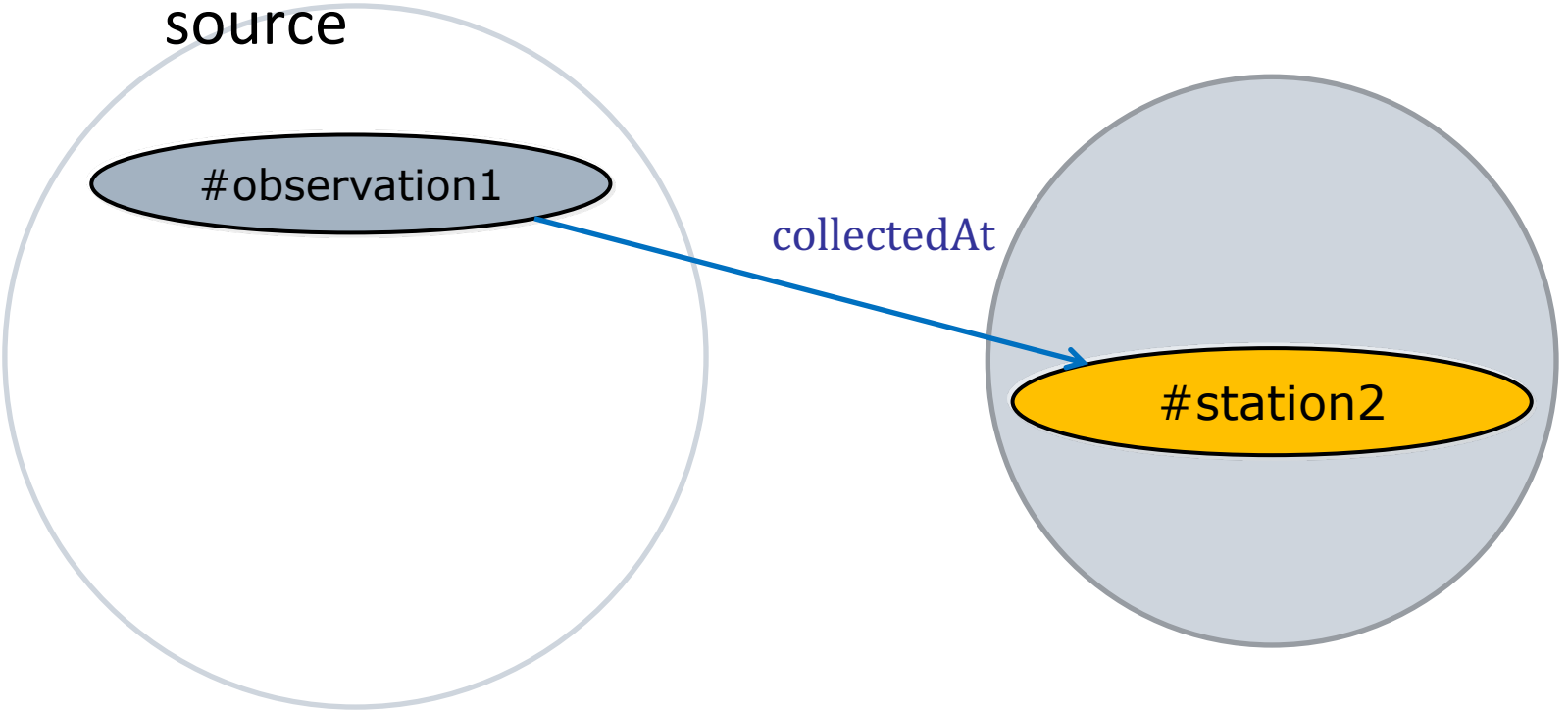
- General methods & frameworks
- ARIF in SDB
- ARIF & voovle

Summary

Next Work

Linking scientific data

- Goal
  - **technically speaking,** linking scientific data means set RDF Links among data resources, which perhaps are in the same data source or not in the same data source



Linking scientific data

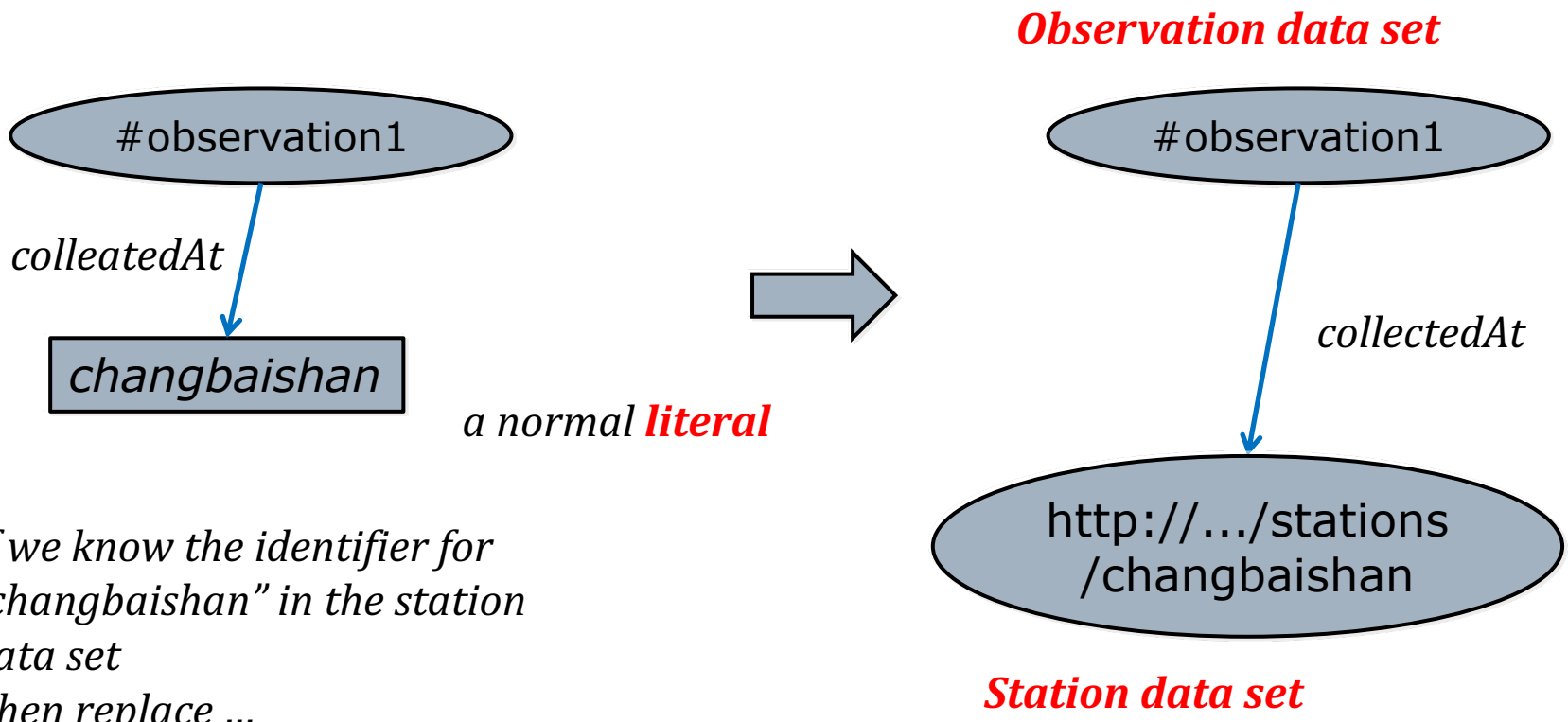
- General methods for generating links
  - Setting RDF Links Manually
  - Auto-generating RDF Links
    1. **identification schemata based**
    2. **similarity of entities based**

1. Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data-the story so far." *International Journal on Semantic Web and Information Systems (IJSWIS)* 5.3 (2009): 1-22.

Linking scientific data

Method 1: identification schemata based method

- If the source and the target data sets already both support one identification schema, the implicit relationship between entities in both data sets can easily be made explicit as RDF links.



If we know the identifier for "changbaishan" in the station data set
Then replace ...

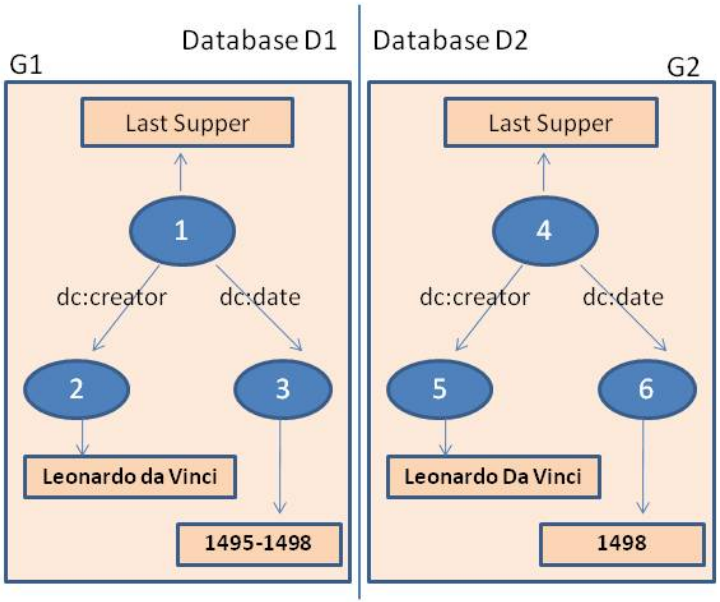
Linking scientific data

- **Method 1: identification schemata based method**
  - Simple and popular
  - in the publication domain there are ISBN and ISSN numbers
  - in life science, various accepted identification schemata exist for genes, molecules, and chemical substances

Linking scientific data

- **Method 2: similarity of entities based method**
  - complex
  - related work
    - Database community: record linkage and duplicate
    - Semantic Web Community: ontology matching

Linking scientific data



similarity computing

nodes		Sim(ni, nj)
n1	n4	1
n2	n5	0.94
n3	n6	0.44

Generating links between two graphs:

1. *retrieve all nodes of two graphs;*
2. *compute similarity of nodes of shared property one by one;*
3. *Make a combination of all sim(node), get sim(g1, g2);*
4. *If sim(g1,g2)>T, then create links;*

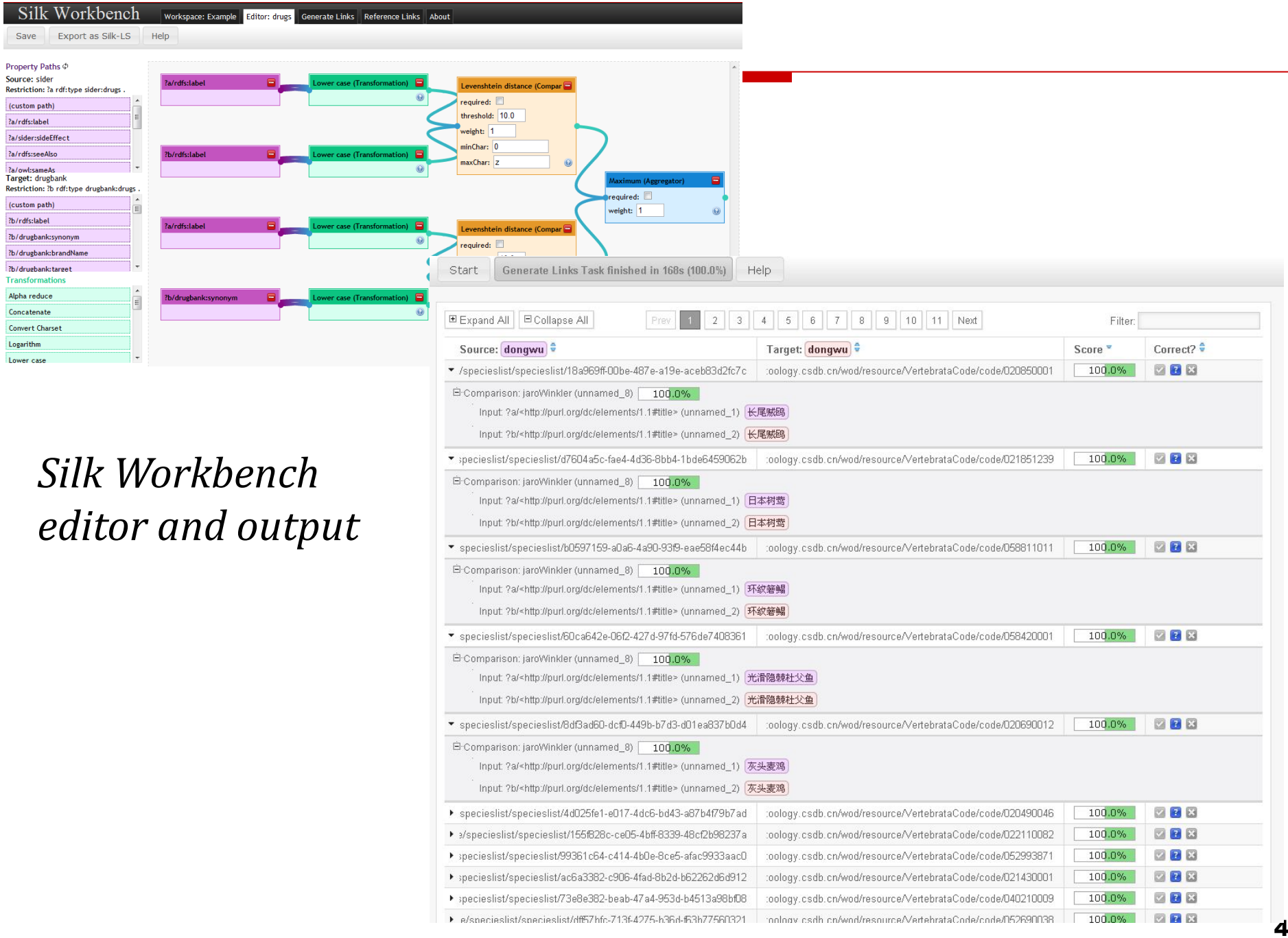
sim(g1, g2)
=sim(n1,n4)+sim(s2,n5)+sim(n3,n6)/3
=0.79

Linking scientific data

- Link Discovery Frameworks
  - SILK
    - a link discovery framework using a declarative language for searching relationships between various datasets.
  - LIMES
    - LIMES implements novel time-efficient approaches for link discovery in metric spaces.
  - RDF-AI
    - RDF-AI focuses on the integration of RDF datasets. Providing modules for pre-processing, matching, fusing, interlinking and postprocessing
  - LinQuer
    - LinQuer is a system for generating SQL queries for semantic link discovery over relational data.
  - LDIF
    - Intergrates tools including LDSpider、R2R、SILK and Sieve to help application developers with these tasks.

Linking scientific data

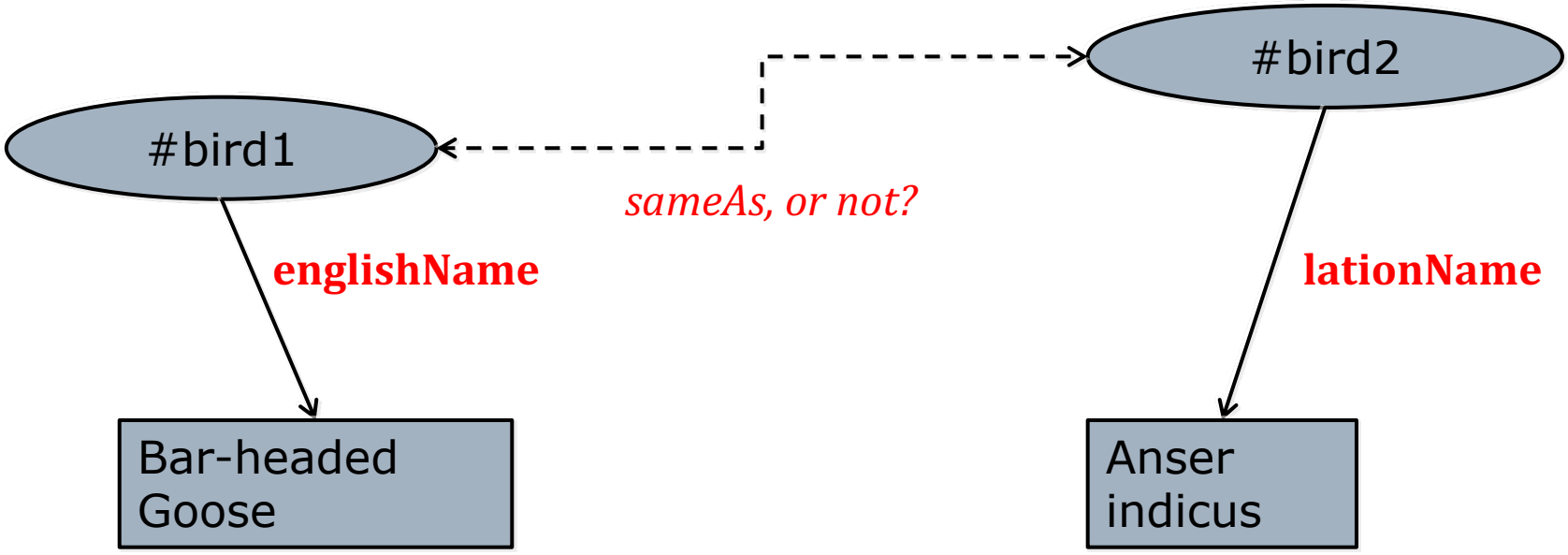
- *SILK as an example*
  - *Silk LSL*
    - Flexible, declarative language for specifying linkage rules
  - *Silk Link Discovery Engine*
    - responsible for loading the instances from the data sources as well as generating the links based on the user-provided Link Specifications.
  - *Silk Workbench*
    - a web application which guides the user through the process of interlinking different data sources



Silk Workbench editor and output

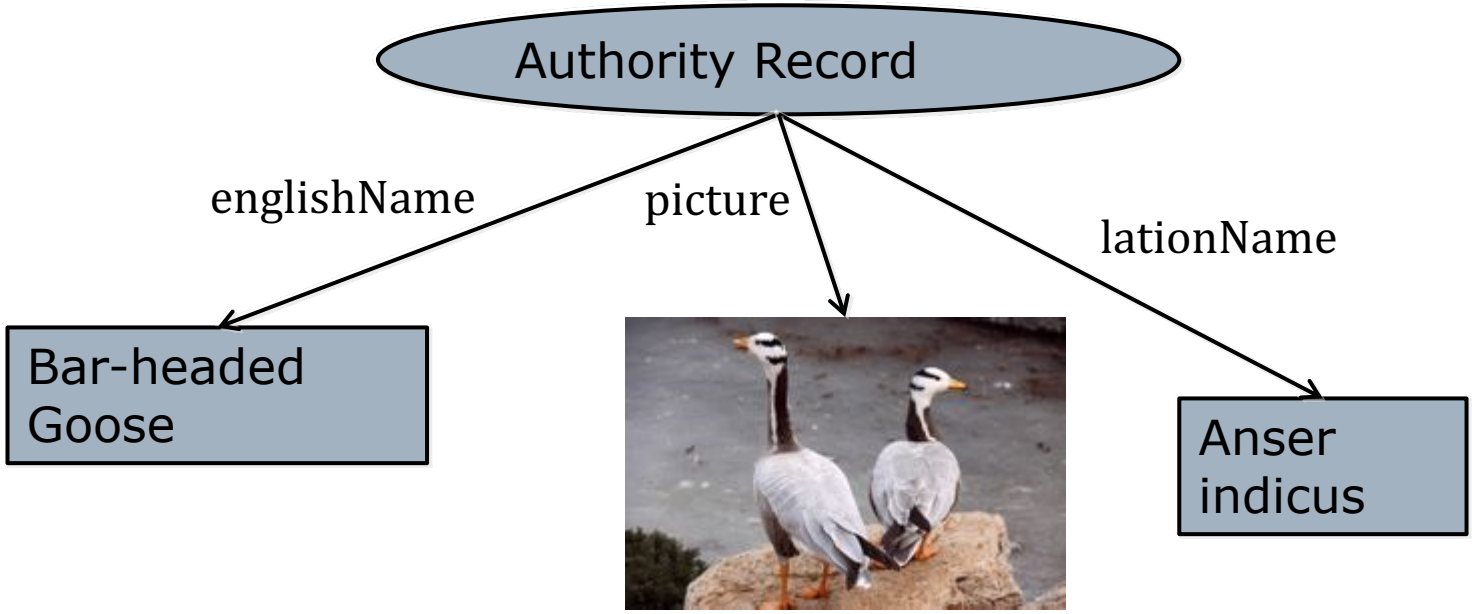
Linking scientific data

- Why these frameworks are not enough for SDB?
- Reason#1
  - Current methods compare two data entities by their shared properties. However, it is not common that two distributed data entities have shared properties



Linking scientific data

- An au'thority record with complete properties can help to find meaningful links



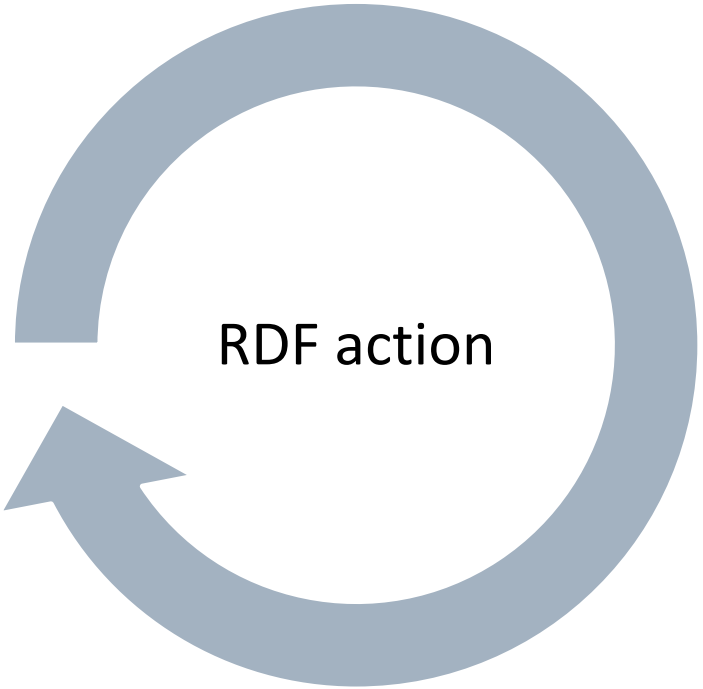
- LORD: A linked open dataset which contains authority records is called LORD (Linked Open Reference Database)
- Some existing LOD datasets (GeoNames) can be used as LORD
- Frameworks like SILK do not support this method

Linking scientific data

- Why these frameworks are not enough for SDB?
- Reason#2
  - The link discovery process is really a long and complex process
    - multiple tasks, paths, steps
  - The process should be able to be planned and a good pipeline mechanism is required
- All frameworks aim at one-step-matching, users have to do some additional work between two steps
  - for example, copy or merge RDF resources

Linking scientific data

- We developed ARIF (Another Resource Interlinking Framework)
  - All tasks are represented as RDF actions
  - ARIF defines 5 kinds of RDF actions

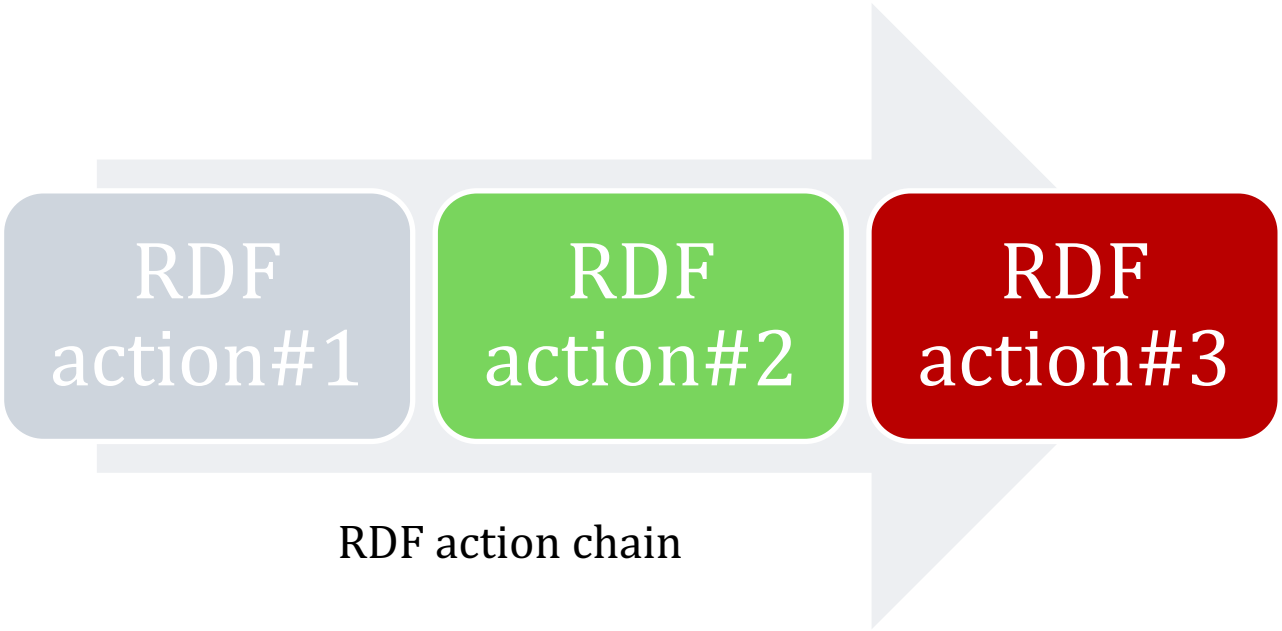


- *RDF dumping*
- *RDF construction*
- *RDF matching*
- *RDF identification*
- *RDF substitution*

LORD based identification

Linking scientific data

- ARIF (Another Resource Interlinking Framework)
  - RDF actions can be chained with I/O parameters
  - RADL : An RDF action can be represented in RADL language
  - ARIFEngine: Task loader and runner



RDF action chain

Linking scientific data

```
<#step1> a :ConstructionStep;
        rdfs:comment "从分类文本中提取分类代码";
        #输入数据源
        :input
        (
                [
                        a :QuerySelection;
                        :from <#rdfdb>;
                        :where """
                                ?database
dc:subject ?subjectText.
                                ?database a
csdb:Database.
                        """
                ]
                <#dispatcher>
        );
        #重构规则设置
        :construction
        [
                :output <#extractedSubjects>;
                :template """
                        ?extracted a skos:Concept.
                        ?extracted dc:identifier ?id.
                        ?extracted skos:prefLabel ?title.
                """;
        ];
```

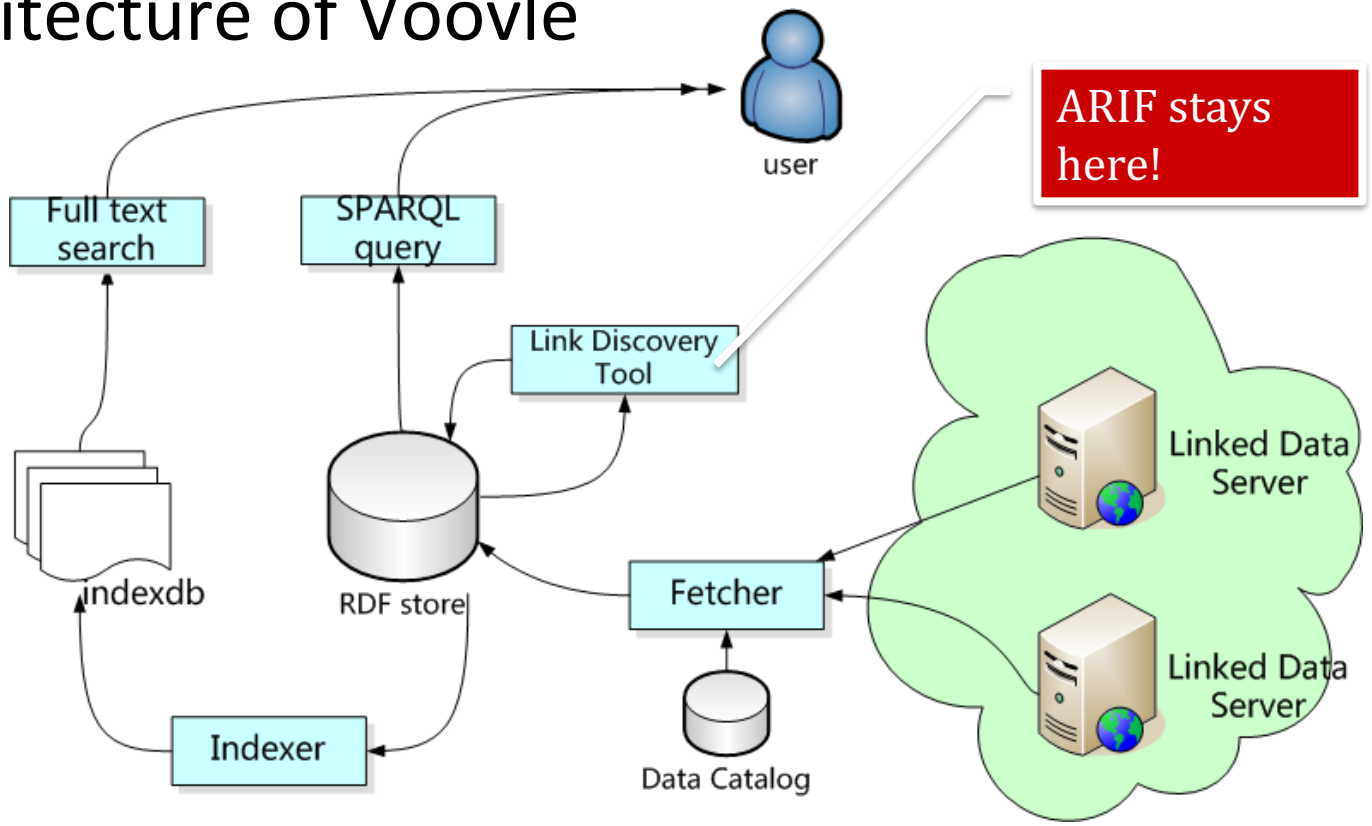
An example of a construction task written in RADL

Searching scientific data

- In SDB, ARIF is implemented and deployed as a part of Voovle
- Voovle is a search engine for SDB
- Voovle offers
  - Keywords-based query service
  - SPARQL (RDF Query) query service

Searching scientific data

- Architecture of Voovle



- Voovle collects all RDF data from distributed data sources into a whole RDF store.

Searching scientific data

- In SDB, ARIF has
  - discovered 1552 links between two plant collection databases
  - discovered 4934 links between two animal databases
  - discovered 244 links between Qinghailake ecological database and animal database
  - Need more experiences…

Searching scientific data

- When displaying a record in Voovle, discovered linked data will be listed

Description of an observation station.

observation towers in the station

Background

Publishing scientific data

Linking scientific data

Summary

Next Work

Summary(1)

- Publishing scientific data
  - We chose RDF and Linked Data as publishing standard
  - We publish datasets, files, and records via VisualDB

Summary(2)

- Linking scientific data
  - We think that the popular link discovery frameworks lack supports of third-party authority databases, and lack pipeline mechanisms.
  - We present ARIF (another resource interlinking framework) and integrate it in Voovle

Next Work

- New challenges
  - The stability of data publishing tool, especially on dealing with a large amount of data
  - How to evaluate the effectiveness of link discovery framework?
  - Improve ARIF to deal with large sized RDF databases

Thanks for your attention!

SHEN Zhihong

bluejoe@cnic.cn