中德开放出版研讨班

Sino-German Training Workshop on Open Access

# 以关联数据发布科学数据

## *Publishing Scientific Data as Linked Data*

沈志宏, 何少鹏, 侯艳飞
Zhihong SHEN, Shaopeng HE, Yanfei HOU
中国科学院计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences
2014/03/13

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

# *Outline*

- 背景
  background
- 为什么选择关联数据
  why Linked Data?
- 关键问题与发布方案
  publishing methods and steps
- 发布工具与应用系统
  publishing tool and Linked Data applications
- 总结与展望
  future work on data publication

# background - Scientific DataBase project

- data sharing project SDB
- 中国科学院信息化专项“科学数据库项目 (SDB)”
- A long-term mission funded by CAS(Chinese Academy of Sciences)
- started in 1986
- mission
  - Collected multi-discipline research data（收集不同学科领域的科研数据）
  - promoted data sharing（促进数据共享）
- *data from research, data for research*

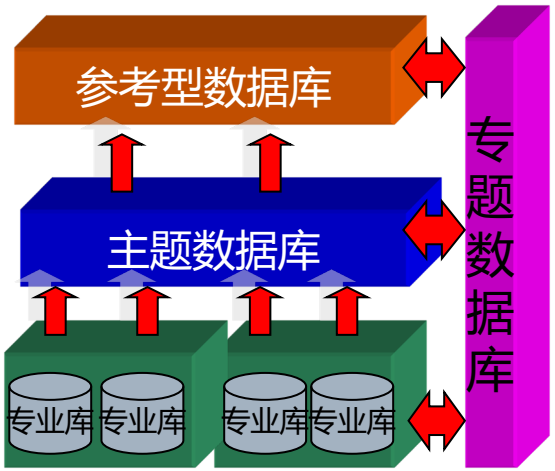

*http://www.csdb.cn*

# *background - Scientific DataBase project*

- During the period of the eleventh Five-Year-Plan of CAS (2006-2010)
- “十一五”期间
  - about 61 CAS institutes involved
  - 有61个中国科学院研究所参与
  - Over 200TB data available for open access and download
  - 超过 200TB 的数据已经开放使用和下载

# 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*




化合物和植物物种2个参考型库

化学、材料、空间、天文、遥感、人地系统、动物、微生物等8个主题库

聚变、青海湖、冰雪冻土、生态功能区划等4个专题库

土壤、海洋、地球化学、指纹、语料、光学、基因组、蛋白组等37个专业库

# background - Scientific Databases

- 51 distributed web sites were set up for databases, following a set of construction standards（数据规范、建站规范、服务规范）


http://www.space.csdb.cn/

http://www.chemdb.csdb.cn

http://www.cryosphere.csdb.cn

http://www.fusion.csdb.cn/

http://www.chemcpd.csdb.cn/

http://www.data.ac.cn

http://www.qinghailake.csdb.cn

http://www.lakesci.csdb.cn

http://www.ocdb.csdb.cn

http://www.soil.csdb.cn

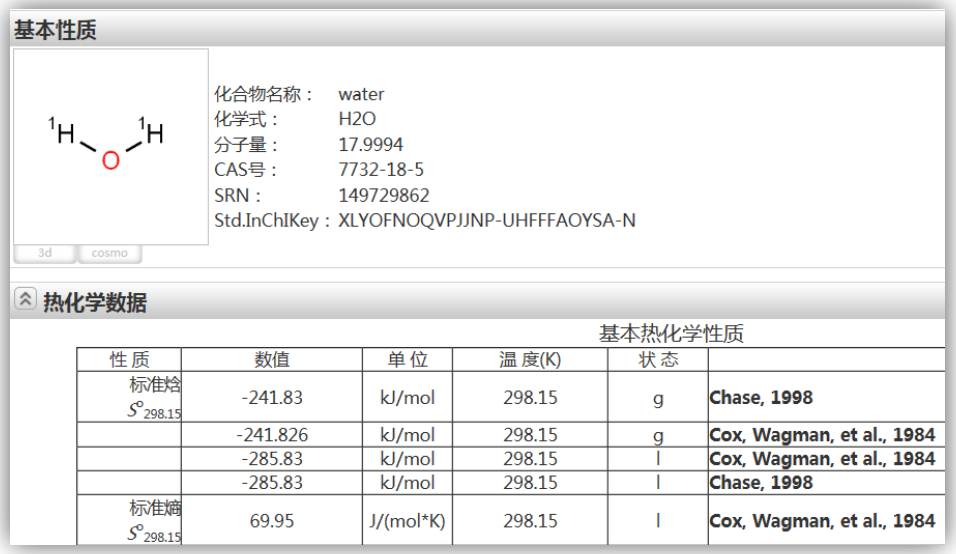
...

# background - Scientific Databases

- Examples

$H_20$的性质数据
attributes of $H_2O$→

*from* http://www.chemdb.csdb.cn




| 性质 | 数值 | 单位 | 温度(K) | 状态 | |
|---|---|---|---|---|---|
| 标准焓 $S^{\circ}_{298.15}$ | -241.83 | kJ/mol | 298.15 | g | **Chase, 1998** |
| | -241.826 | kJ/mol | 298.15 | g | **Cox, Wagman, et al., 1984** |
| | -285.83 | kJ/mol | 298.15 | l | **Cox, Wagman, et al., 1984** |
| | -285.83 | kJ/mol | 298.15 | l | **Chase, 1998** |
| 标准熵 $S^{\circ}_{298.15}$ | 69.95 | J/(mol*K) | 298.15 | l | **Cox, Wagman, et al., 1984** |

鸟的GPS信息
GPS locations of a bird→

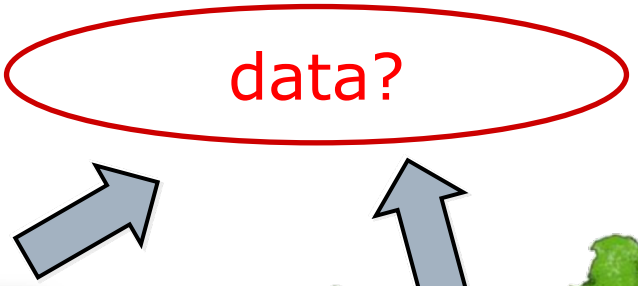
*from* http://qinghailake.csdb.cn

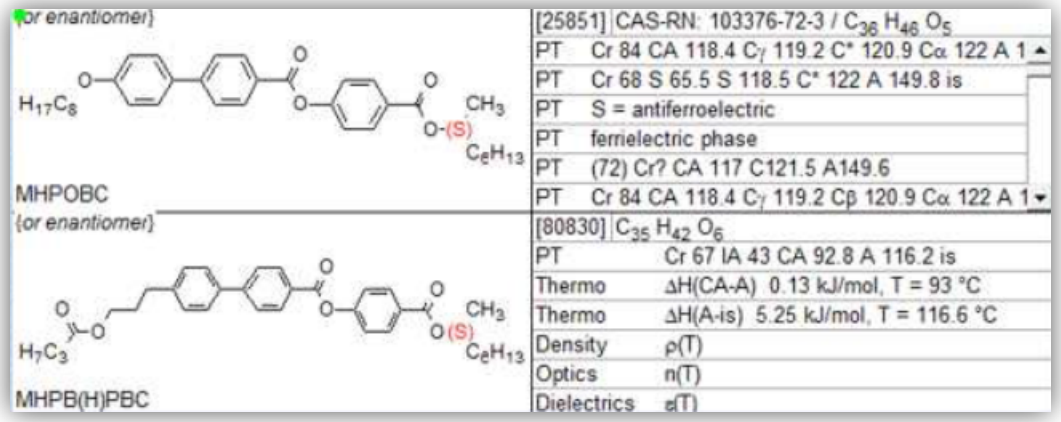
| | |
|---|---|
| Id | 465347 |
| Obs | 11 |
| 编号 | BH07_67580 |
| Ptt | 67580 |
| Date | 2007-06-23 |
| Record Id | LATEST ARGOS LOCS |
| Time | 15:16:04 |
| Days Ago | 652 |
| Hrs Ago | 15644.87 |
| Latitude | 36.132 |
| Longitude | 98.805 |
| Lc94 | LB |
| Nmess | 2 |
| Voltage | |
| Days Dply | 89 |

# *background – problems*

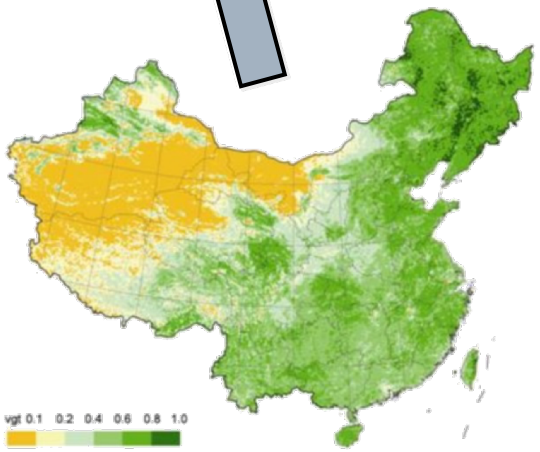
- The big problem is that …
- Scientific data can be browsed easily on the Web, but can not be easily accessed and understood by computer programs

- 数据的展示很精彩，但…
  应用程序无法获取到数据！




HTML presentation



HTML presentation

# *background – motivation*

- Thus we need a standard method to make distributed scientific data
  - *both Accessible*
  - *and Understandable*

- Such a standard method includes
  - *an unified data representation format*（统一的数据描述格式）
  - *an unified open access mechanism*（统一的数据访问机制）

# *Outline*

- 背景
  background
- 为什么选择关联数据
  why Linked Data?
- 关键问题与发布方案
  publishing methods and steps
- 发布工具与应用系统
  publishing tool and Linked Data applications
- 总结与展望
  future work on data publication

# *why Linked Data?*

- main features of scientific data（科学数据的特征）
  - variety of data formats（学科领域差异带来的格式多样性）
    - Storage forms（存储格式的多样性）
      - structured, semi-structured, non-structured
    - Property value types（属性值类型的多样性）
      - text, date, GPS location, images…
  - in-born interlinked（普遍存在的关联性）
    - gene and disease（基因与疾病）
    - Materials and chemical compounds（材料与化学）
    - researchers, locations, objects of ecological observation records
    - 生态考察的观测人员、地理位置、对象信息

# *why Linked Data?*

- requirements on data representation format（科学数据发布格式的要求）
  1. able to represent data in different forms and formats（包容格式多样性）
  2. able to describe the links among data（能表达关联）
  3. ensure consistent understanding of names (class names, property names…)保证领域内名字含义的一致性
  4. flexible schema, easy to extend, adding a new property is very simple（无模式：属性易扩展）
  5. makes full use of existing description tools, esp. XML（对现有描述格式的继承，特别是XML格式）
     - CML、MathML、SMILES, ...
     - DarwinCore

# *why Linked Data?*

- We chose RDF as data representation format
  - RDF: **Resource Description Framework**

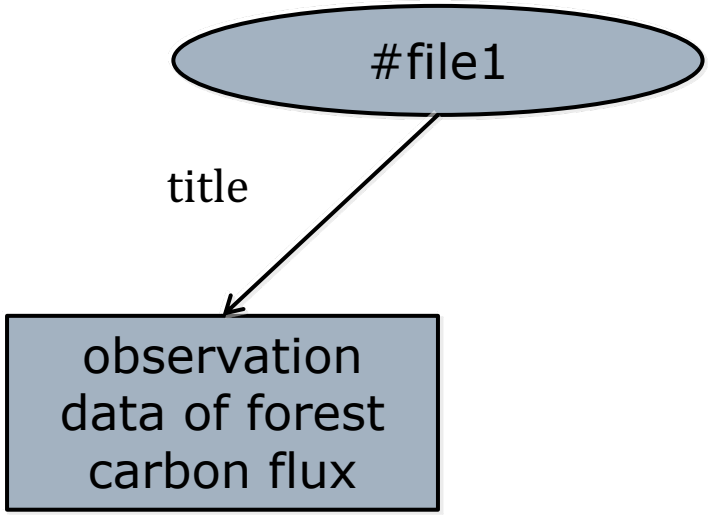
    *http://www.w3.org/**RDF***

  - RDF describes resources in terms of simple properties and property values（属性：属性值）
  - it is based upon the idea of *making statements about resources* (in particular web resources) in the form of subject-predicate-object expressions
  - 声明：主-谓-宾 三元组
  - e.g

| <#file1> | <#title> | “observation data of forest carbon flux” |
|---|---|---|
| subject | predicate | object |
| resource | property | value |

# *why Linked Data?*

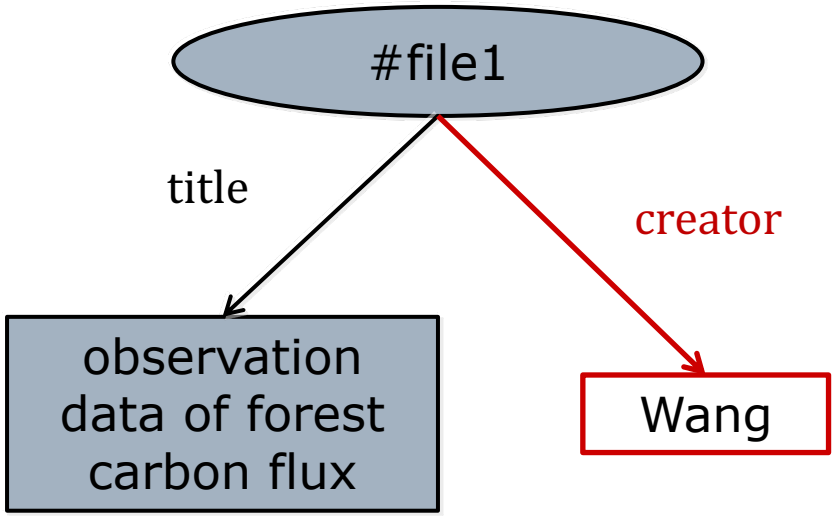
- RDF statements about a resource are often represented as a graph（RDF图：节点、弧）
  - nodes: representing the resources, and their properties values.
  - arcs: representing properties of resources




<#file1> <#title> “observation data of forest carbon flux”

# *why Linked Data?*

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



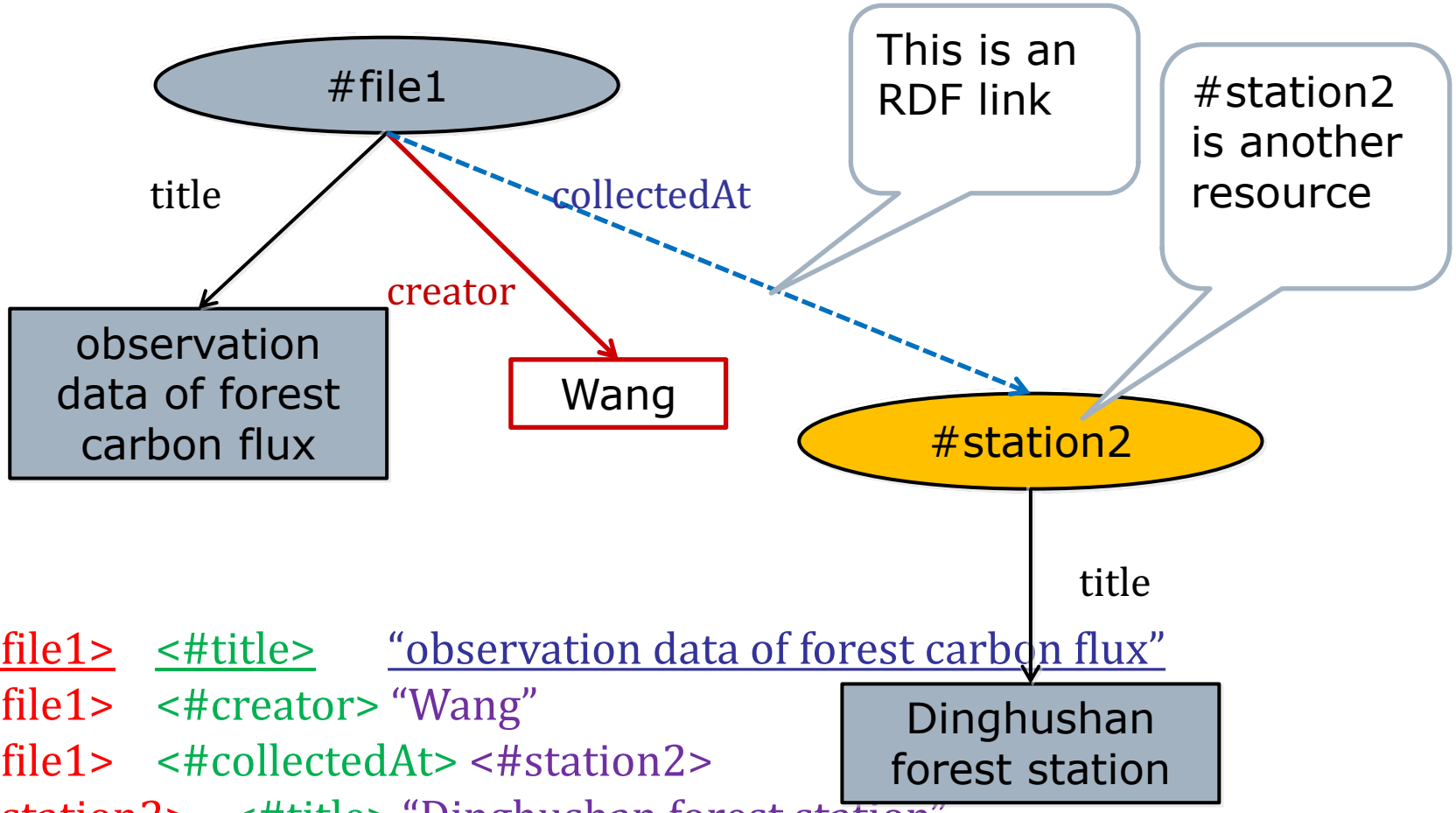


<#file1> <#title> “observation data of forest carbon flux”

<#file1> <#creator> “Wang”

# *why Linked Data?*

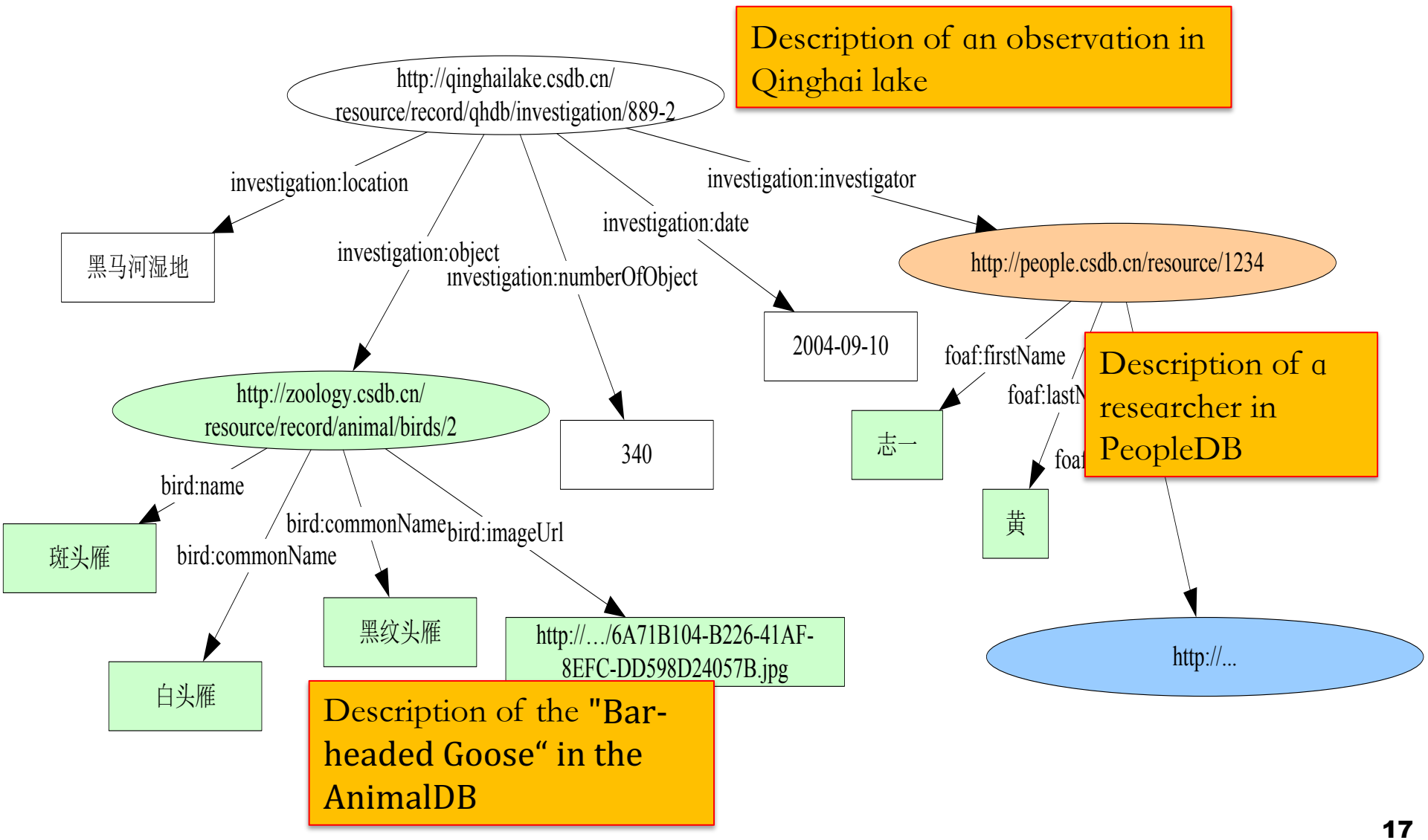
- RDF链接：Most important, RDF data model enables people to set RDF links between data from different sources.




<#file1> <#title> "observation data of forest carbon flux"
<#file1> <#creator> "Wang"
<#file1> <#collectedAt> <#station2>
<#station2> <#title> "Dinghushan forest station"

# why Linked Data?

- More complex RDF data example



# *why Linked Data?*

- After choosing the data format, we chose Linked Data as the open access mechanism of data sources
  - Tim Berners-Lee, the father of World Wide Web, coined the term Linked Data in 2006[1].
  - 万维网之父Tim Berners-Lee在2006年提出Linked Data
  - 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

# *why Linked Data?*

- Four basic principles of Linked Data
  Linked Data 的“四项基本原则”

1. *Use URIs as names for things*
   *采用URIs来命名事物*
2. *Use HTTP URIs so that people can look up (dereference) those names.*
   *采用HTTP URIs以便用户可以查询这些名字*
3. *When someone looks up a URI, provide useful information.*
   *当用户查询URI时，提供有用的信息*
4. *Include links to other URIs so that they can discover more things.*
   *包含到其他资源的链接，以便发现更多的信息*

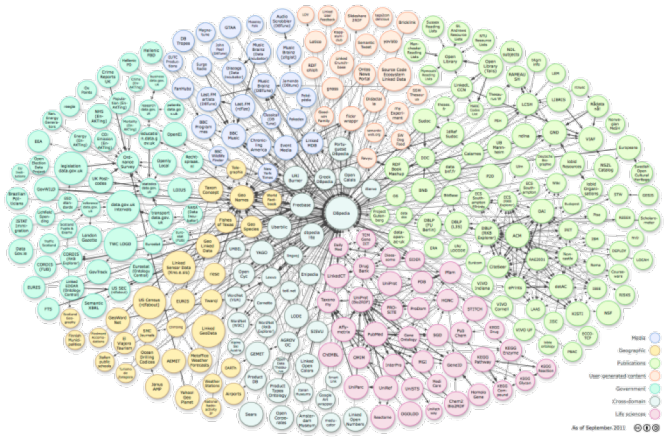
1. T. Berners-Lee, "Design issues: Linked Data," Online at http://www.w3. org/DesignIssues/LinkedData. html, 2006

# *why Linked Data?*

- 关联数据
  - 完全架构于目前的Web体系之上，这对科学数据库来说几乎意味着零升级成本。
  - Linked Data is a standard solution which is totally built on current Web architecture, that means very little cost for upgrading SDB project.

Web of Document

Linked Data

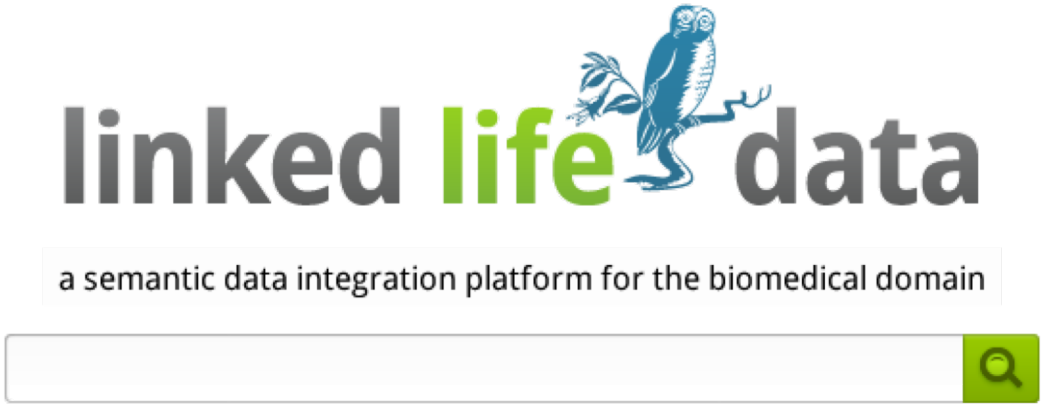


Web of Data

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

# *why Linked Data? - scientific data examples*

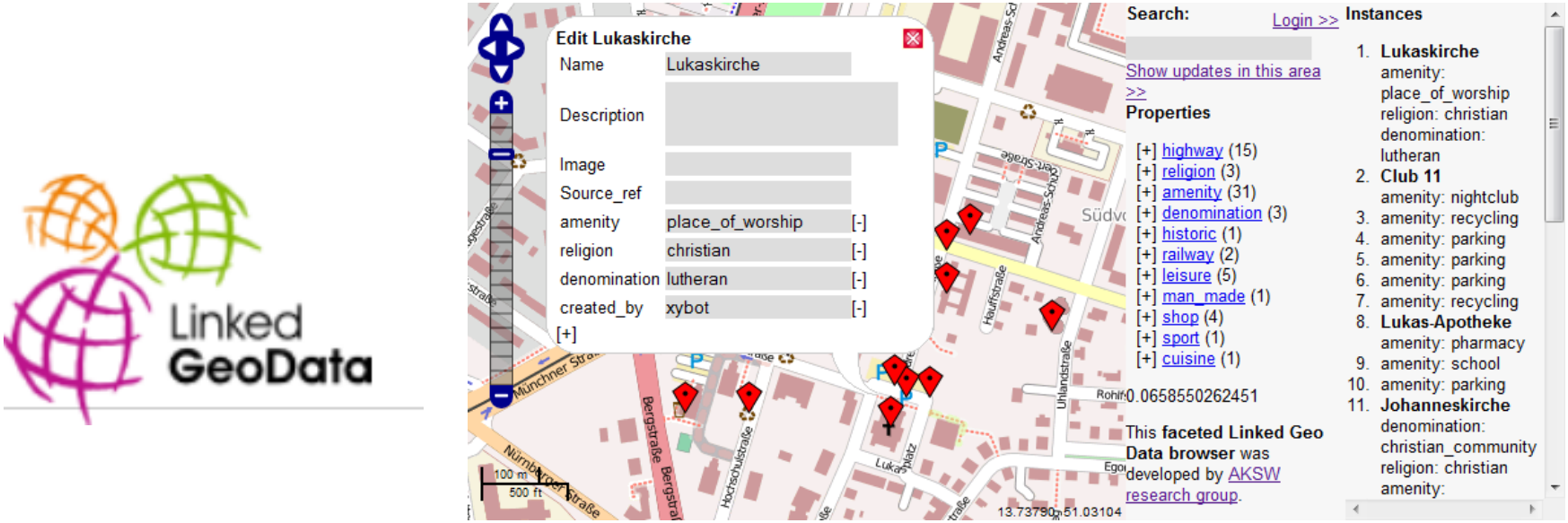
- Linked Life Data
  - Searches and explores over RDF statements from various sources including UniProt, PubMed, EntrezGene and 20 more…
  - 包括UniProt, PubMed, EntrezGene 等20多个数据源，10亿条记录
  - Performs complex SPARQL(RDF query) queries and retrieves more than one billion RDF resources.




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.

# *why Linked Data? - scientific data examples*

- LinkedGeoData
  - uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles.
  - 采用OpenStreetMap 收集的信息并依据Linked Data原则将其发布为可用的RDF知识集

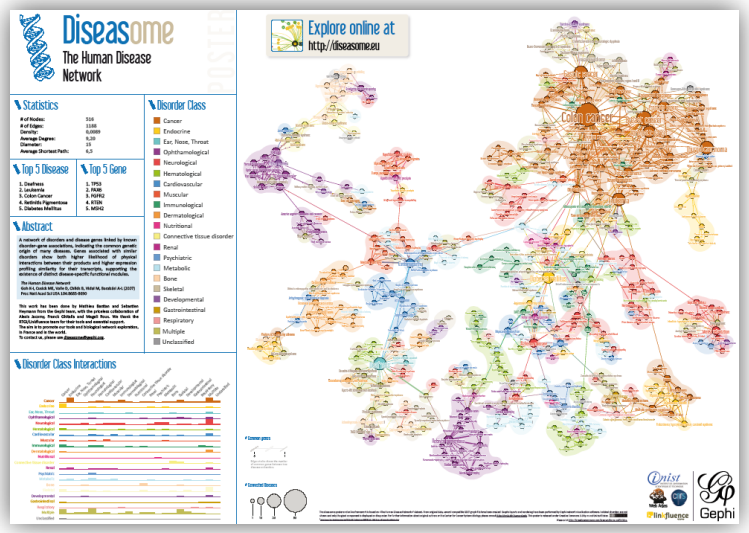




1. Auer, Sören, Jens Lehmann, and Sebastian Hellmann. "Linkedgeodata: Adding a spatial dimension to the web of data." The Semantic Web-ISWC 2009. Springer Berlin Heidelberg, 2009. 731-746.

# why Linked Data? - scientific data examples

- **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.
  - 发布了4,300疾病和疾病基因的关联数据，探索疾病表现和基因之间的关联关系，并绘制血多疾病常见的基因起源




- **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.
  - 发布关于生物序列、家族、物种、物种出险记录、相关数据、以及到geonames, bio2rdf, dbpedia, freebase, umbel的链接

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

# *Outline*

- 背景
  background
- 为什么选择关联数据
  why Linked Data?
- 关键问题与发布方案
  publishing methods and steps
- 发布工具与应用系统
  publishing tool and Linked Data applications
- 总结与展望
  future work on data publication

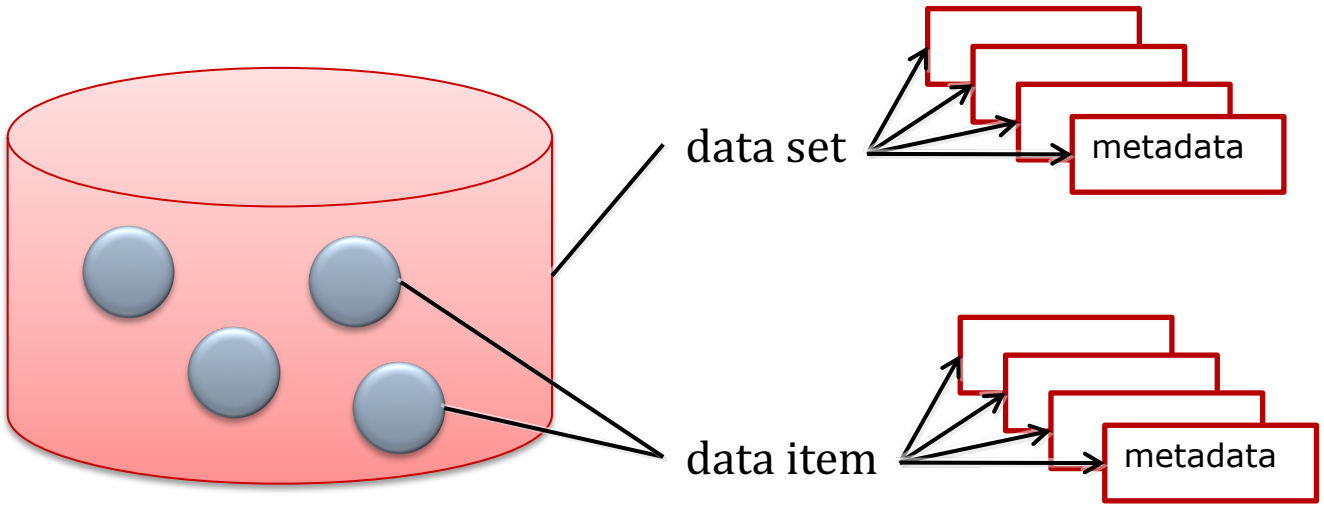
# *publishing methods and steps*

- two issues should be considered in the publishing process
  - Issue #1: data items own little information
    Scientific data contains many files/records which are not easy to be found only by their own property values（科学数据的记录非常多，但是单凭它的属性还很难被发现）
  - Issue #2: RDF data are structured, but most scientific data are unstructured.
    unstructured scientific data (usually exist as data files) are much more than structured scientific data（除了结构化数据，科学数据更多的是文件）

# *publishing methods and steps*

- Issue #1: data items own little information

- We define data in two 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 methods and steps*

- Data set 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/

Elements:

*<dc:title>*
*<dc:subject>*
*<dc:source>*
*<dc:description>*
*<dc:coverage>*
*<dc:creator>*
*<dc:contributor>*
*<dcterms:accrualPeriodicity>*
*<dcterms:rightsHolder>*
*<prism:creationDate>*
*<prism:keyword>*
*<dc:type>*
*<dcterms:rights>*
*<prism:publicationDate>*
*<rdfs:label>*
*<prism:url>*
*<csdb:dqInfo>*
*<csdb:sharePolicy>*
*<csdb:purpose>*

# *publishing methods and steps*

- An example of data set 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 methods and steps*

- Issue #2: RDF data are structured, but most scientific data are unstructured
- We distinguish data items into two classes

**Data records（数据记录）**

- records usually stored in relational databases
- structured

| | A | B | C | | |
|---|---|---|---|---|---|
| 1 | College Enrollment 2005 - 2006 | | | | |
| 2 | | | | | |
| 3 | **Student ID** | **Last Name** | **Initial** | **Age** | **Program** |
| 4 | ST348-245 | Smith | B. | 21 | Drafting |
| 5 | ST348-246 | Wilson | C. | 19 | Science |
| 6 | ST348-247 | Thompson | S. | 18 | Business |
| 7 | ST348-248 | James | D. | 23 | Nursing |
| 8 | ST348-249 | Ramirez | A. | 37 | Science |
| 9 | ST348-250 | Graham | T. | 20 | Arts |
| 10 | ST348-251 | Rosen | O. | 26 | Business |
| 11 | ST348-252 | Hirsch | W. | 22 | Arts |
| 12 | ST348-253 | Russell | E. | 20 | Nursing |
| 13 | ST348-254 | Robitaille | K. | 19 | Drafting |

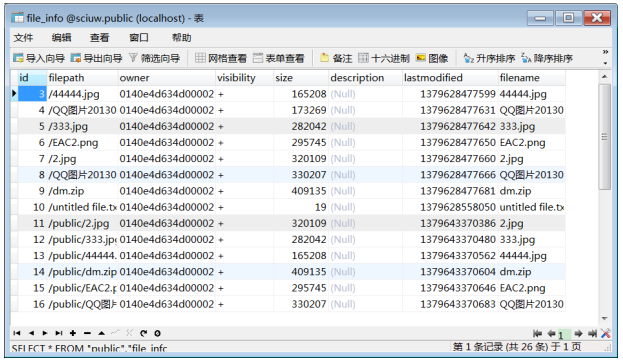
Each row in the table is a data record

**Data files（数据文件）**

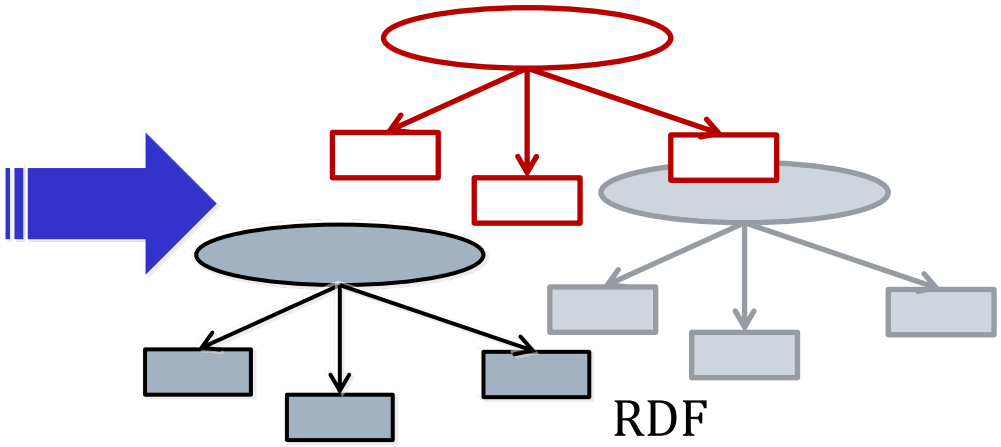
- File content is unstructured
- Metadata is structured

# *publishing methods and steps*

- Publishing data records - rules:
  1. *each table is mapped to a data set*
     *一个表对应一个数据集*
  2. *each record is mapped to an RDF resource*
     *一条记录对应一个RDF资源*
  3. *each field is mapped to an RDF property*
     *每个字段对应RDF的一个属性*
  4. *field values are mapped to RDF property values*
     *字段值对应RDF的属性值*
  5. *foreign key relationships are mapped to RDF links*
     *外键关联映射成RDF链接*




data records



RDF

# *publishing methods and steps*

- Publishing data files – rules:
  1. *Each file is assigned two URLs, one for file download, the other for file metadata*
     *两个URL：文件下载+文件描述*
  2. *When users access download URL, provides file contents in binary-stream (Non-RDF)*
     *发布文件的二进制内容*
  3. *When users access file description URL, provides mixed description in RDF format*
     *发布文件的描述信息*
  4. *file metadata includes file attributes and extracted metadata*
     1. *file attributes（文件的属性信息）*
        - *filename, size, creation time*
        - *often retrieved from file systems*
     2. *extracted metadata from content（自动抽取的元数据）*
        - *FITS、HDF4、JPG、NetCDF、PowerPoint、Visio、Word*

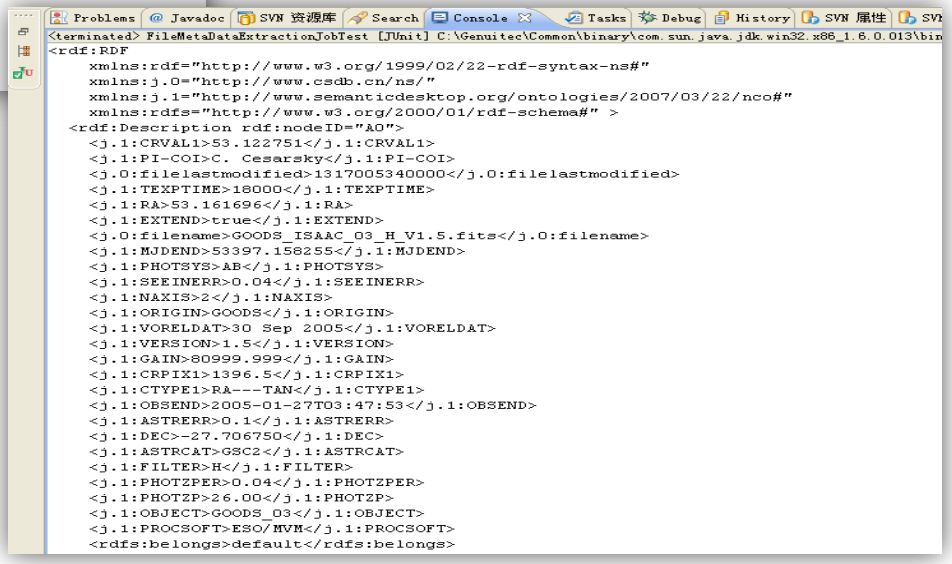
# *publishing methods and steps*

- Examples of extracted metadata

```
<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



# *Outline*

- 背景
  background
- 为什么选择关联数据
  why Linked Data?
- 关键问题与发布方案
  publishing methods and steps
- 发布工具与应用系统
  publishing tool and Linked Data applications
- 总结与展望
  future work on data publication

# *publishing tool - VisualDB*

- 科学数据发布工具VisualDB
- In SDB project, we developed VisualDB help data owners to publish scientific data both as HTML pages and Linked Data on the Web
- 完全配置化发布，无需编程
- VisualDB allows users define publishing rules in Web interfaces, without programming


*Visual Scientific Data Management and Publishing Tool*

# *publishing tool - VisualDB*

- In SDB, VisualDB has been deployed in 37 institutes
  - 4.1 billion records（41亿条记录）
  - 26 million files（2600万个文件）
  - are published as Linked Data

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | College Enrollment 2005 - 2006 | | | | |
| 2 | | | | | |
| 3 | Student ID | Last Name | Initial | Age | Program |
| 4 | ST348-245 | Smith | B. | 21 | Drafting |
| 5 | ST348-246 | Wilson | C. | 19 | Science |
| 6 | ST348-247 | Thompson | S. | 18 | Business |
| 7 | ST348-248 | James | D. | 23 | Nursing |
| 8 | ST348-249 | Ramirez | A. | 37 | Science |
| 9 | ST348-250 | Graham | T. | 20 | Arts |
| 10 | ST348-251 | Rosen | O. | 26 | Business |
| 11 | ST348-252 | Hirsch | W. | 22 | Arts |
| 12 | ST348-253 | Russell | E. | 20 | Nursing |
| 13 | ST348-254 | Robitaille | K. | 19 | Drafting |

Each row in the table is a data record

4.1 billion records

26 million files

# *publishing tool - VisualDB*

- Sample URIs

*request for a html page*

http://www.zoology.csdb.cn/wod/page/VertebrataCode/code/020370019

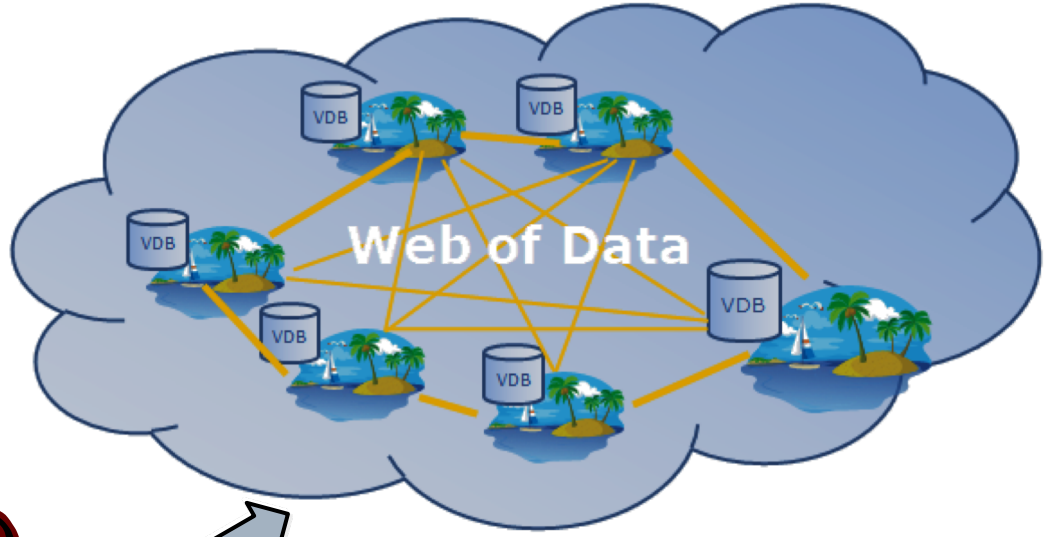

*request for an RDF graph*

http://www.zoology.csdb.cn/wod/data/VertebrataCode/code/020370019

```
<rdf:RDF
  xmlns:dc="http://purl.org/dc/elements/1.1#"
  xmlns:code="http://www.zoology.csdb.cn/wod/resource/vocab/VertebrataCode/code/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:j.0="http://www.zoology.csdb.cn/wod/resource/vocab/VertebrataCode/"
  xmlns:dataset="http://voovle.csdb.cn/ns/dataset#" >
 <rdf:Description
rdf:about="http://www.zoology.csdb.cn/wod/resource/VertebrataCode/code/020370019">
  <code:family>Anatidae</code:family>
  <code:genus>Anser</code:genus>
  <code:cfamily>鸭科</code:cfamily>
  <code:cgenus>雁属</code:cgenus>
  <code:corder>雁形目</code:corder>
  <code:csubfamily>null</code:csubfamily>
  <dataset:belongsEntity
rdf:resource="http://www.zoology.csdb.cn/wod/resource/vocab/VertebrataCode/code"/>
  <rdf:type
rdf:resource="http://www.zoology.csdb.cn/wod/resource/vocab/VertebrataCode/code"/>
  <code:species>indicus</code:species>
  <code:order>ANSERIFORMES</code:order>
  <code:subfamily>null</code:subfamily>
  <code:subspecies>null</code:subspecies>
  <code:author>(Latham , 1790)</code:author>
  <code:code>020370019</code:code>
  <code:cname>斑头雁</code:cname>
  <dataset:belongs
rdf:resource="http://www.zoology.csdb.cn/wod/resource/vocab/VertebrataCode"/>
  <dc:title>斑头雁</dc:title>
 </rdf:Description>
</rdf:RDF>
```

# *Linked Data application - Voovle*

- 科学数据搜索引擎Voovle
- We developed Voovle as a search engine for scientific data published as linked data.




*How to search?*

Voovle

# *Linked Data application - Voovle*

- Voovle offers free text search by inputting keywords（文本检索）
- Voovle also offers semantic query by submitting a SPARQL query（支持SPARQL查询）

```
SELECT distinct (?s) WHERE
{
?s ?p '斑头雁'
}
```



time is:1400 ms

| No | .0 |
|---|---|
| 1 | http://www.zoology.csdb.cn/wod/resource/Animalnames/name/AE8132F8-4FF5-40BC-BCBF-326D15796110 |
| 2 | http://www.zoology.csdb.cn/wod/resource/cnAtlas/tableTaxa/8A40049F-D2C0-455F-8409-A54F5B419DED |
| 3 | http://www.zoology.csdb.cn/wod/resource/cnfauna/tableTaxa/1478B56D-6609-4A58-8305-CFB75621AB0A |
| 4 | http://www.zoology.csdb.cn/wod/resource/specieslist/specieslist/27db1a96-b1e4-40b2-bd50-bf5ace063bb1 |
| 5 | http://www.zoology.csdb.cn/wod/resource/VertebrataCode/code/020370019 |
| 6 | http://www.qinghailake.csdb.cn/wod/resource/cn.csdb.qhnew/circlelakeVirussample/1813 |
| 7 | http://www.qinghailake.csdb.cn/wod/resource/cn.csdb.qhnew/circlelakeVirussample/1812 |
| 8 | http://www.qinghailake.csdb.cn/wod/resource/cn.csdb.qhnew/circlelakeVirussample/1826 |
| 9 | http://www.qinghailake.csdb.cn/wod/resource/cn.csdb.qhnew/birds/2 |

[1] Apache Lucene - Welcome to Apache Lucene. Online at http://lucene.apache.org/

# *Linked Data application - Voovle*

Voovle

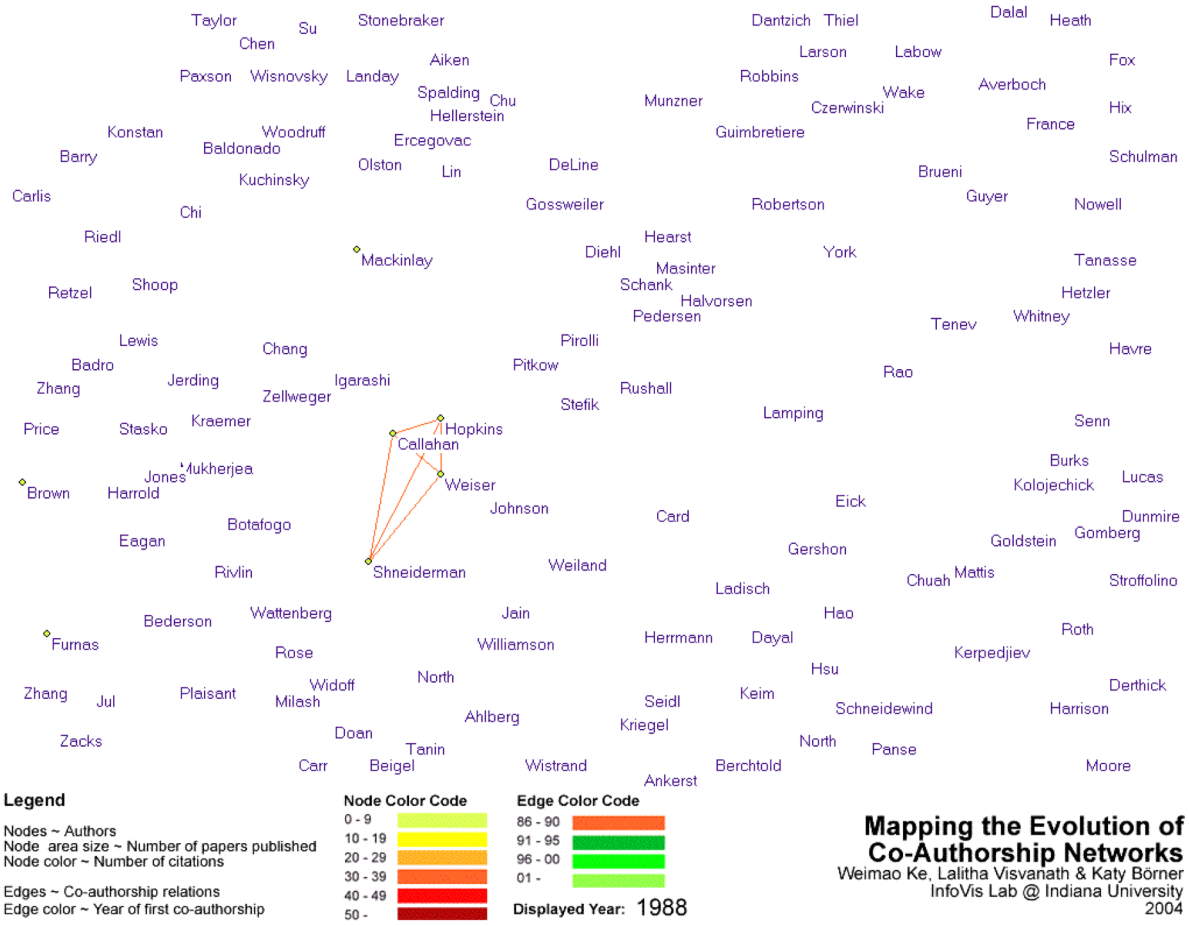

*search engine for SDB*
*http://voovle.csdb.cn*
*124 datasets, 5.64 million records*


*search engine for CERN observation data*
*http://cern.voovle.csdb.cn*


*search engine for Scientific Data Grid*
*http://datagrid.voovle.csdb.cn*

# *Linked Data application – link discovery tool*

- Upon RDF store built by Voovle, we developed tools to find useful links among scientific data.
- 跨数据集的关联发现



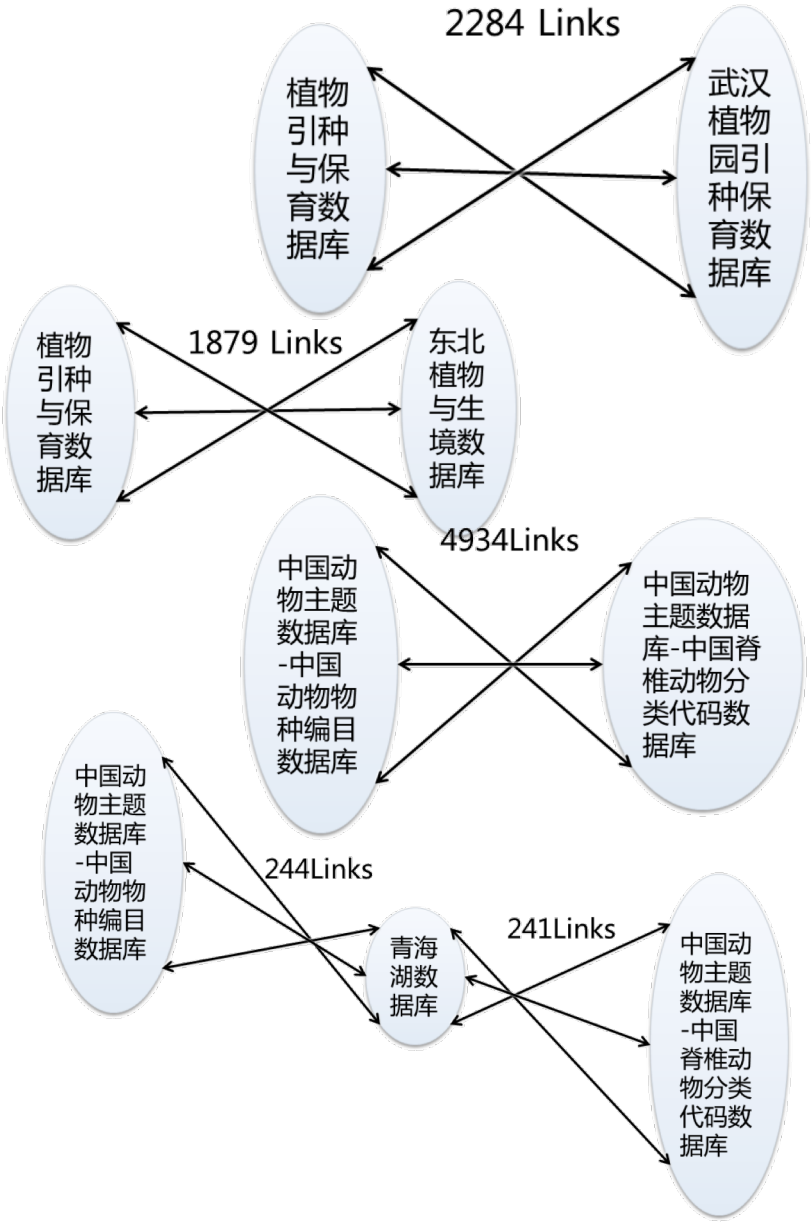
# *Linked Data application – link discovery tool*

- ARIF is such a tool based on graph similarity（图形相似度）
- In ARIF, all RDF graph operations involved in the link discovery process are defined as different "RDF actions".
- ARIF归纳了关联发现过程中的所有RDF图形操作，并定义为RDF action。


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

# *Linked Data application – link discovery tool*

- 部分关联发现效果
- experiments and results with ARIF
  - discovered 2284 links between two plant collection databases
  - discovered 4934 links between two animal databases
  - discovered 244 links between Qinghailake ecological observation database and animal database
  - trying more experiments…



# *Outline*

- 背景
  background
- 为什么选择关联数据
  why Linked Data?
- 关键问题与发布方案
  publishing methods and steps
- 发布工具与应用系统
  publishing tool and Linked Data applications
- 总结与展望
  future work on data publication

# *conclusions*

- 我们发布科学数据以促进数据共享
  - We published scientific data to promote data sharing in Chinese Academy of Sciences
- 我们选择RDF作为科学数据的发布格式
  - We chose RDF as data representation format for scientific data
- 我们选择关联数据作为科学数据开放访问机制
  - We chose Linked Data as open access mechanism of scientific data

# *new problems*

- no guarantee of quality
  数据质量没有保证

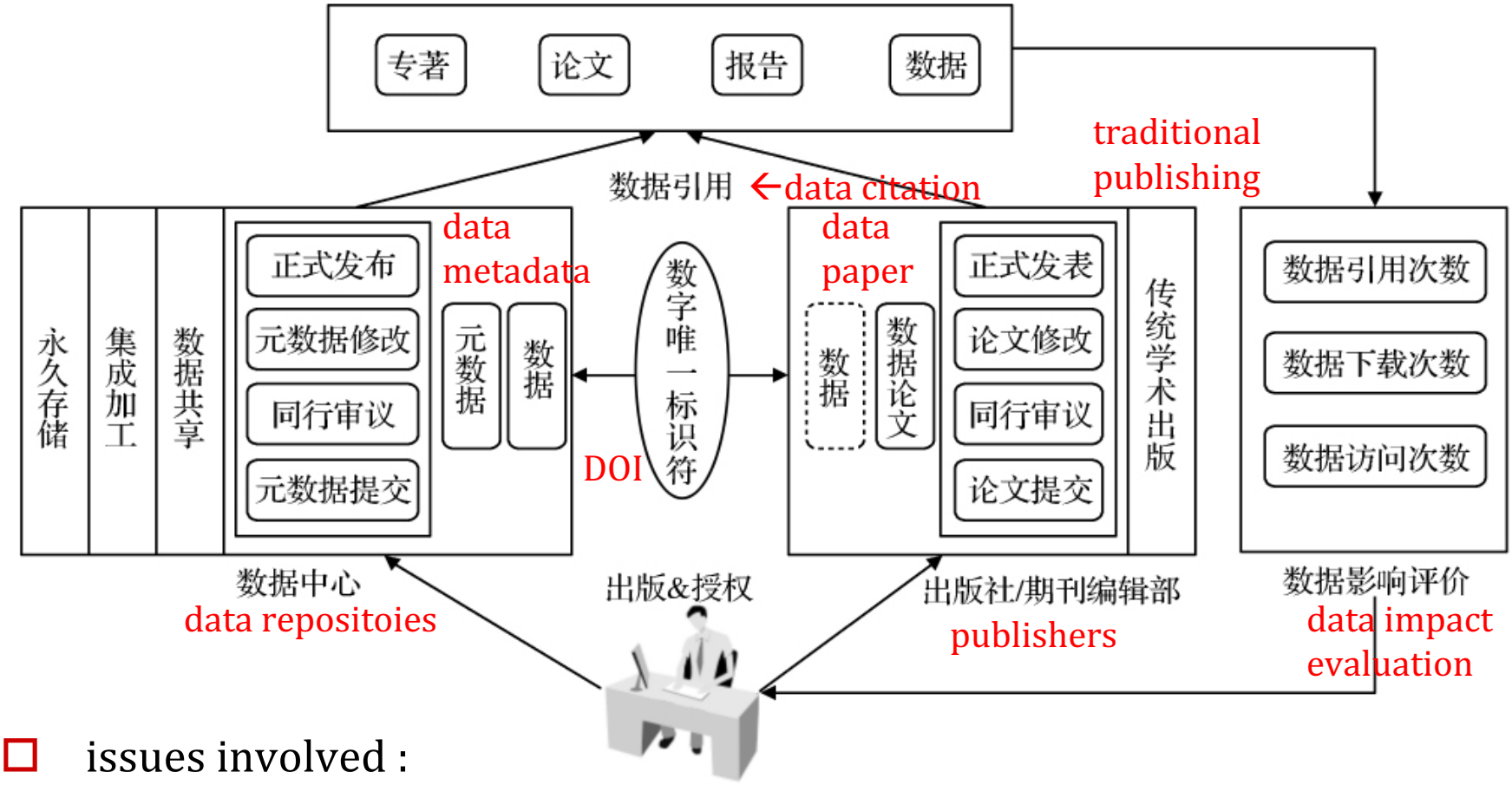
- unstable data access URLs
  数据URL不稳定

- unable to evaluate the values of data
  数据的学术价值无法评估

# *possible answers?*

- no guarantee of quality
  数据质量没有保证
  *—peer reviews on data?*
- unstable data access URLs
  数据URL不稳定
  *—data repositories provide data storage and preservation?*
- unable to evaluate the values of data.
  数据的学术价值无法评估
  *—data DOIs enable data citation?*


# *big picture of Data Publishing(or Publication)*

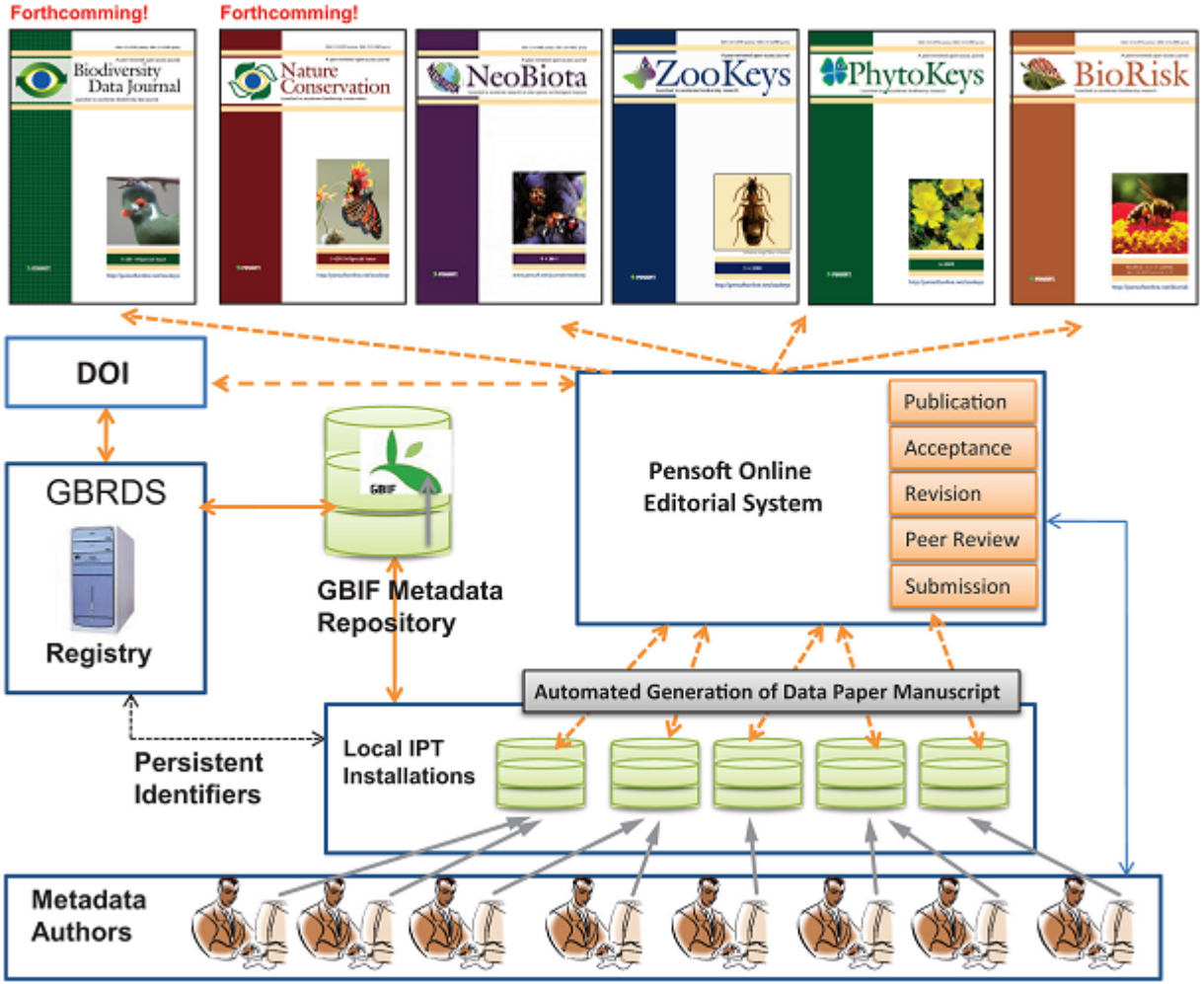



- issues involved :
  - metadata management/data management/identifier/peer review
  - long term storage/data citation/data impact evaluation
- roles involved: data repository/ publisher/ journal/ author

1. 吴立宗等, "科学数据出版现状及其体系框架,"遥感技术与应用, 2013

# *similar framework for Data Publishing*

the GBIF/Pensoft workflow of data publishing and automated generation of Data Paper

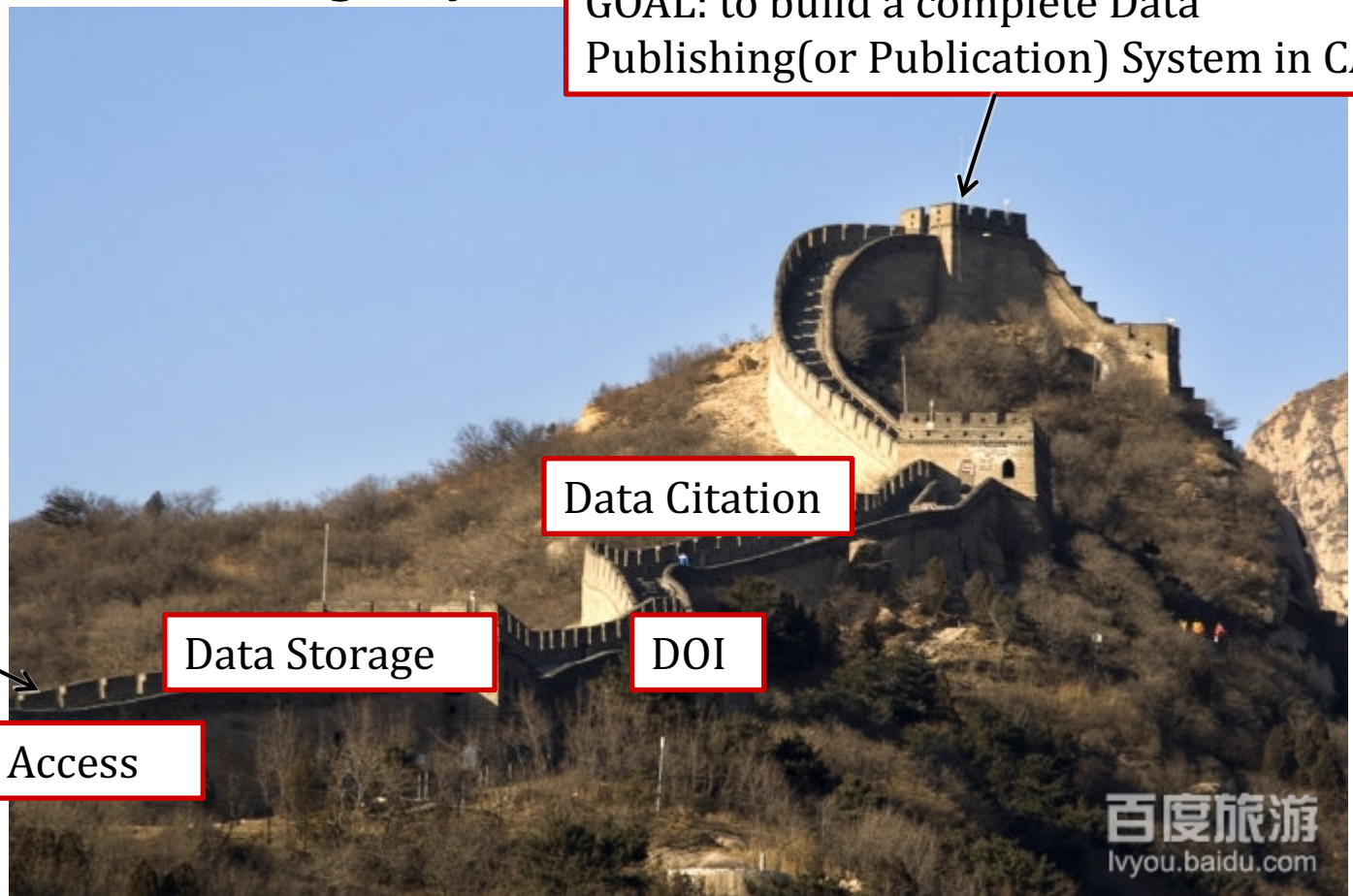



- important issues involved in data publication:
  - Metadata
  - Identifier
  - markup
  - link
  - Archiving
  - migration
  - Exchange
  - rights
  - …


1. Lyubomir Penev etl. Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Implemented by Pensoft Publishers,26th of May 2011

# *Keep walking*

- 千里之行，始于足下
- Our goal is to build a complete data publishing(or publication) system based on CNIC cyber infrastructure, but we still have a long way to



*Thank you for your attention!*

*请提宝贵意见！*

报告人：沈志宏

Zhihong SHEN

中国科学院 计算机网络信息中心

bluejoe@cnic.cn

*http://www.csdb.cn*