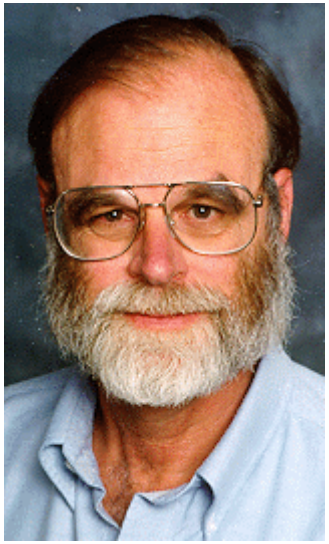
A part of the report "Scientific Research in the Age of Big Data"

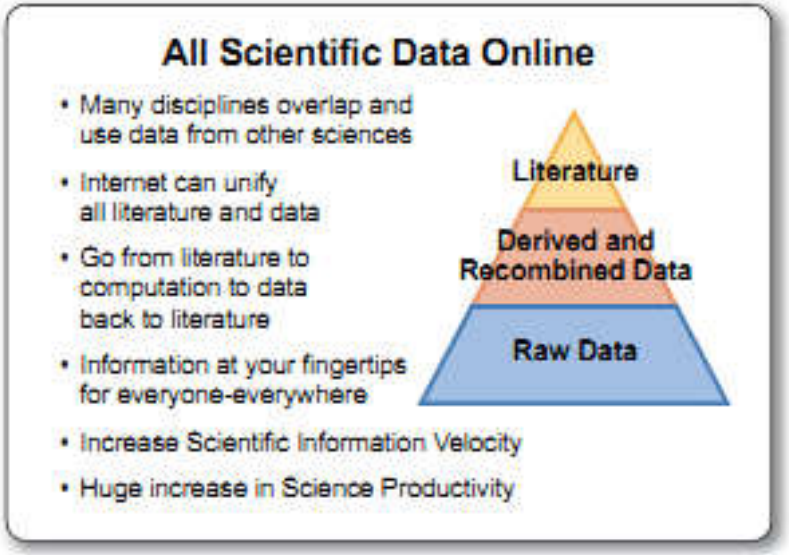
Interoperation between Scientific Data and Literature : An overview

Shen Zhihong
Computer Network Information Center, CAS
2012/10/25



Jim Gray
Microsoft Research's
eScience Group.

http://research.microsoft.com/en-us/um/people/gray/



... the Internet can do more than just make available the full text of research papers. *In principle, it can unify all the scientific data with all the literature to create a world in which the data and the literature interoperate with each other.* You can be reading a paper by someone and then go off and look at their original data. You can even redo their analysis. Or you can be looking at some data and then go off and find out all the literature about this data. Such a capability will increase the "information velocity" of the sciences and will improve the scientific productivity of researchers. And I believe that this would be a very good development!

1. The fourth paradigm: data-intensive scientific discovery. USA: Microsoft Research 2009

Interoperation

- Interoperation
  - customer services effectively combining multiple resources and domains[1]
- Interoperation may use following pattern, methods[2]:
  - Connectors
  - Adapters
  - Converters
  - Simulators
  - Bridges
  - Combination

1. Gio Wiederhold, "Glossary"; in Intelligent Integration of Information, Kluwer Academic Publishers, Boston MA, July 1996, pages 193--203
2. http://en.wikipedia.org/wiki/Interoperation

Outline

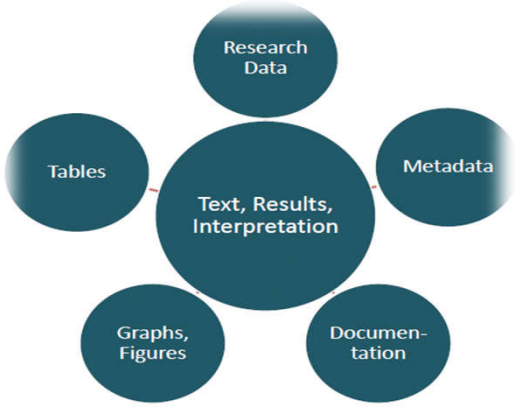
- This paper introduces current state of the interoperation in three main directions:
  1. Data publication and citation
     publishing and citing scientific data like papers – is it an adapter pattern?
  2. Semantic Publishing
     publishing actionable data in articles – is it a combination pattern?
  3. Integrated services
     linking data and literature via integrated search and exploration services – is it a bridge pattern?

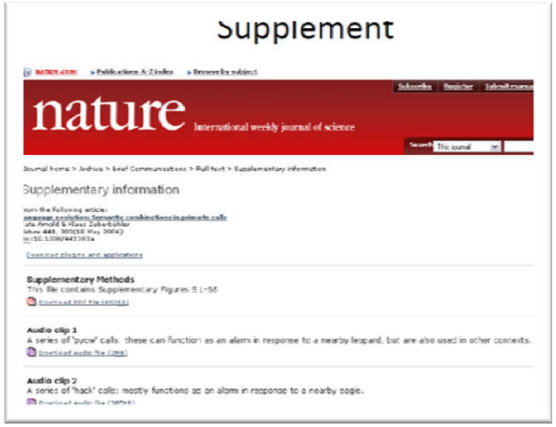
The 1st form of interoperation

- Data publication and citation
  - *publishing and citing scientific data like papers*

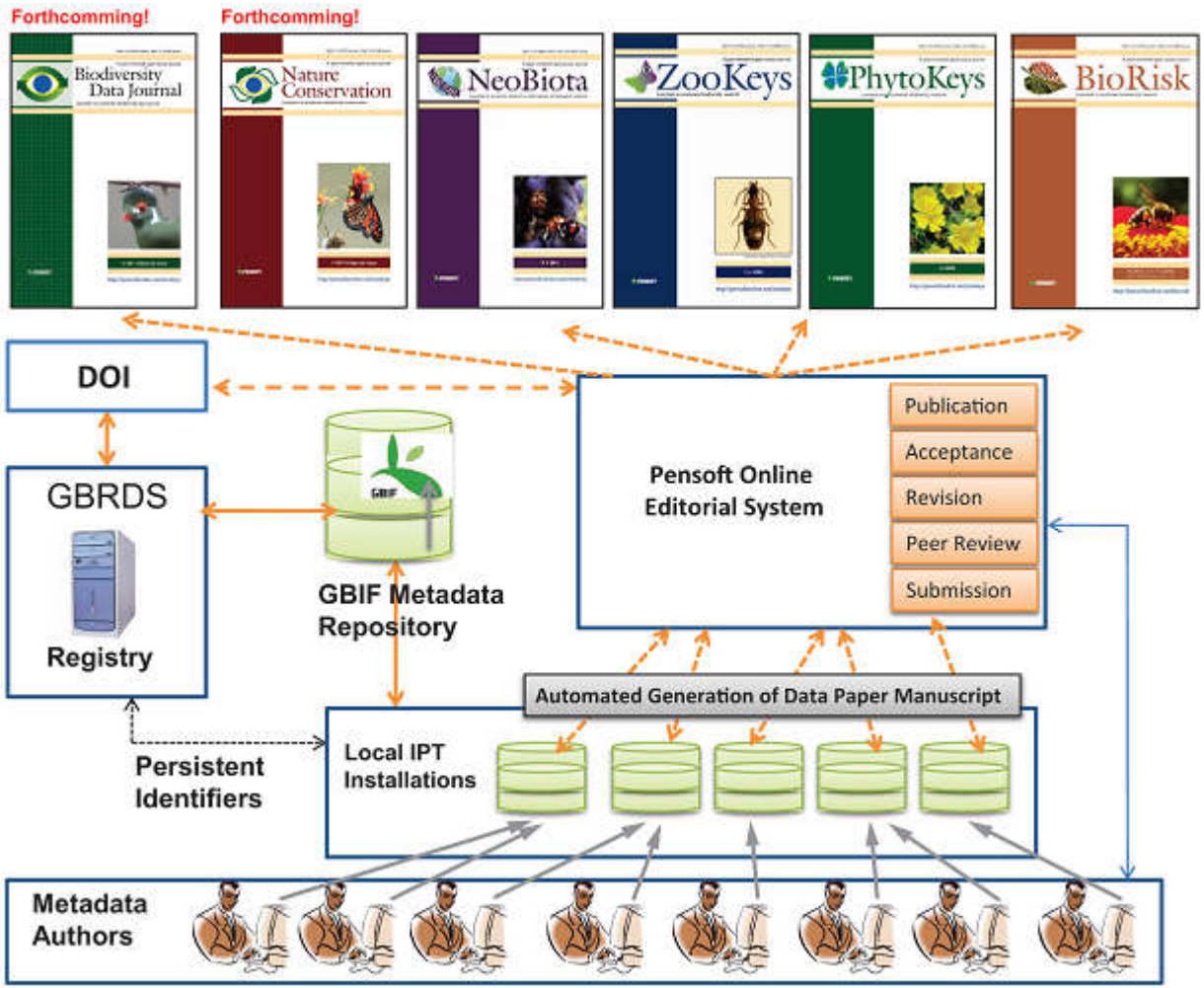
1. Data publication and citation

- Data publication is operated in different forms
- Publishing data as…
  - supplementary materials of articles
    - Sünje Dallmeier-Tiessen. Research Data Publishing. 2010
  - strategic Integration of Article Content
    - Supplement Materials, Linda Beebe, 2011
  - Data Journals (and overlay journals?)
    - **Earth System Science Data**
    - **Journal of Physical and Chemical Reference Data (AIP)**
    - **Journal of Chemical and Engineering Data (ASC)**
    - **Atomic Data and Nuclear Data Tables (Elsevier)**





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



- Important issues invovled in data publication:
  - Metadata
  - Identifier
  - markup
  - link
  - Archiving
  - migration
  - Exchange
  - rights
  - …

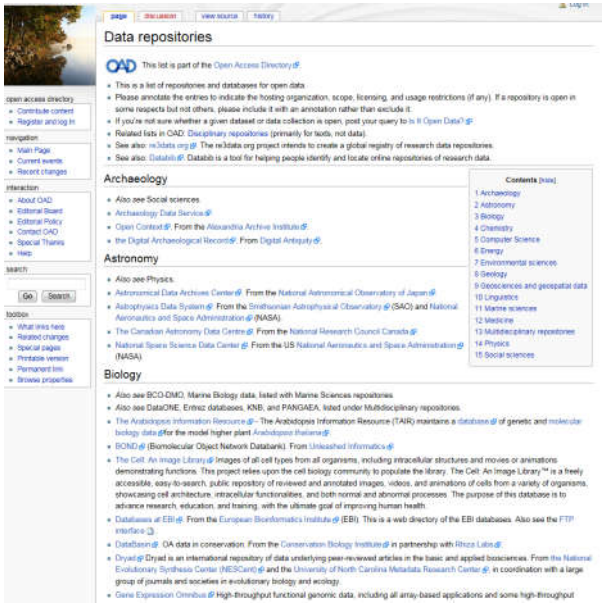
1. Lyubomir Penev etl. Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Implemented by Pensoft Publishers,26th of May 2011

1. Data publication and citation

- Depositing data in repositories
- data repositories (public databases, data warehouses, data hosting centers)
  - are subject- or institution-oriented infrastructures, usually based at large national or international institutions.
  - provide data storage and preservation according to widely accepted standards, and provide free access to their data holdings for anyone to use and re-use under the minimum requirement of attribution

1. Data publication and citation

- Open data repositories
  - An example is Dryad
  - http://datadryad.org/
  - an international repository of data underlying peer-reviewed articles in the basic and applied biosciences
  - By Aug 30, 2012, 1949 data packages and 5145 data files with articles in 147 journals



1. http://oad.simmons.edu/oadwiki/Data_repositorie

1. Data publication and citation

DataCite

http://datacite.org

- global consortium carried by local institutions
- focused on improving the scholarly infrastructure around datasets and other non-textual information
- focused on working with data centres and organisations that hold data
- Providing standards, workflows and best-practice
- Initially, but not exclusivly based on the DOI system

- Founded December 1st 2009 in London

1. Data publication and citation

- DataCite in 2012
- Over 1,300,000 DOI names registered so far
- DataCite Metadata schema published (in cooperation with all members) http://schema.datacite.org
- DataCite MetadataStore
- http://search.datacite.org
- OAI Harvester
- http://oai.datacite.org
- Content negotiation
- http://data.datacite.org

1. Data publication and citation

CODATA task group data citation

http://www.codata.org/taskgroups/TGdatacitation/index.html

Approved at CODATA GA 2010 in South Africa

Wide representation from different stakeholder(data centers, scientists, funders, libraries, publisher)

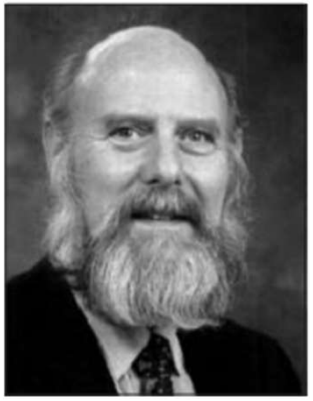
Goals:

- Inventory of existing data citation methods and workflows
- Conduct surveys in the community
- Provide Examples and Recommendations
- Start standardisation proccess

The 2nd form of interoperation

- Semantic Publishing
  - *publishing actionable data in articles*

2. Semantic Publishing



David Shotton

Image Bioinformatics Research Group, Department of Zoology, University of Oxford, Oxford, United Kingdom

- Concept
  - David Shotton, 2009
  - We define the term **semantic publication** to include anything that
    - enhances the meaning of a published journal article
    - facilitates its automated discovery
    - enables its linking to semantically related articles
    - provides access to data within the article in actionable form
    - or facilitates **integration of data between articles**.

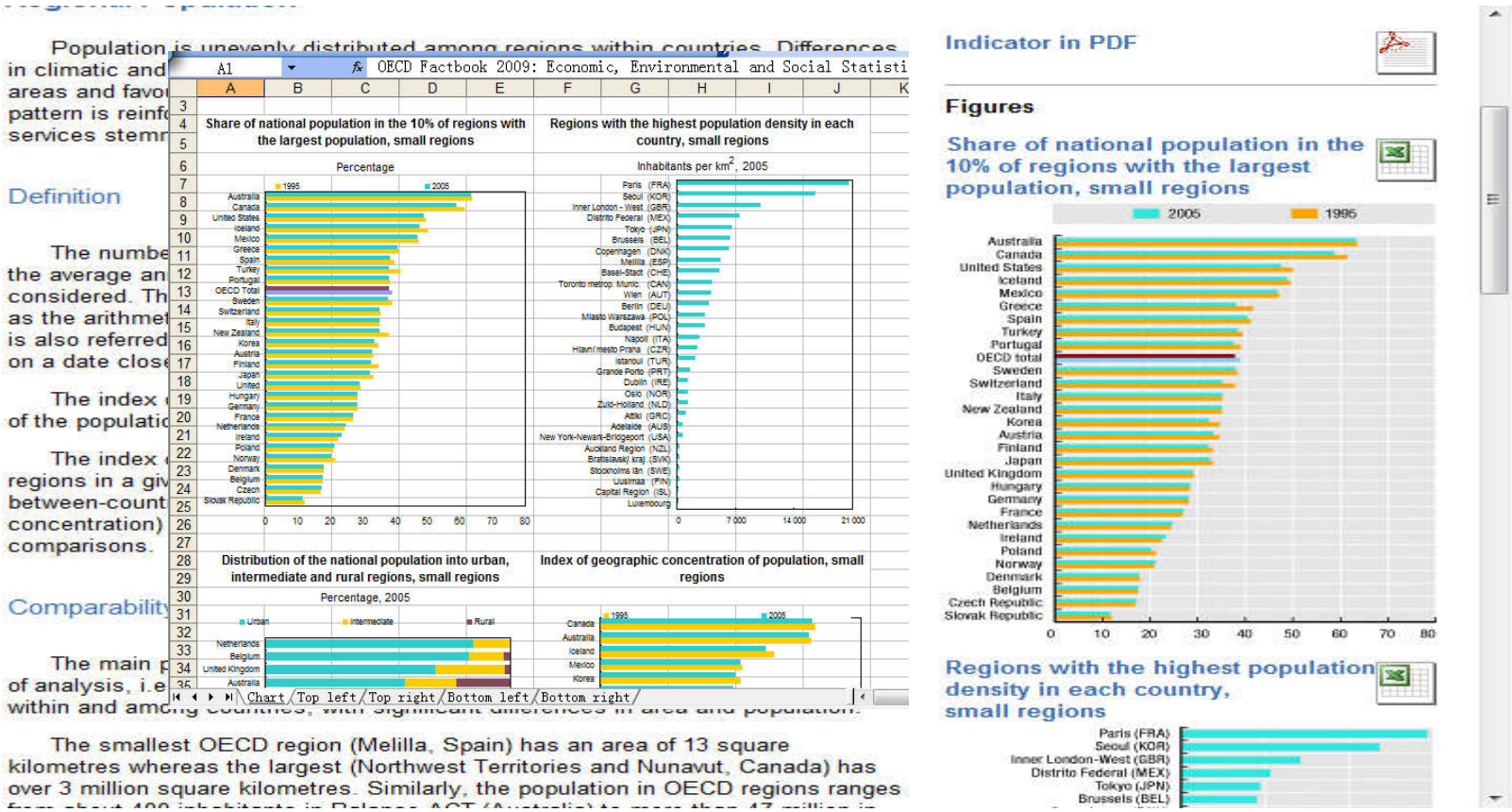
1. Shotton D. Semantic Publishing: the coming revolution in scientific journal publishing. Learned Publishing, 2009 (22) : 85-94.

Shotton introduces some methods of semantic enhancements as examples. Main methods include:

Theme of semantic enhancements	Content of semantic enhancements
Providing Access to Actionable Data	With the cooperation of PLoS who registered new DOIs for us, we then made these spreadsheets downloadable, from the "raw data" links adjacent to the thumbnails for Table 1 and Figure 2 in the enhanced article
Data Fusion with Information from Other Sources	1. Simple geospatial data fusion 2. Geospatial data fusion across multiple publications 3. Mapping leptospirosis study locations in space and Time 4. Serological data fusion across publications
Adding Value to the Text	1. Highlighting of textual terms 2. Links from named entities to external information sources 3. The Supporting Claims Tooltip to permit Citations in Context
Making Information More Accessible	1. Provision of a document summary 2. Study summary 3. Tag cloud. 4. Tag trees 5. Infectious disease ontology terms 6. Document statistics 7. Citation analysis 8. Alternative language abstract 9. Provenance information
User Interactivity	1. Interactive figures 2. Optional re-ordering of the reference list
Provision of New Hyperlinks	1. Links to cited references 2. Hyperlinks to external sites
Machine-Readable Citation Metadata	1. Embedded machine-readable metadata—Use of RDFa 2. Machine-readable self-referencing metadata 3. Machine-readable reference list.

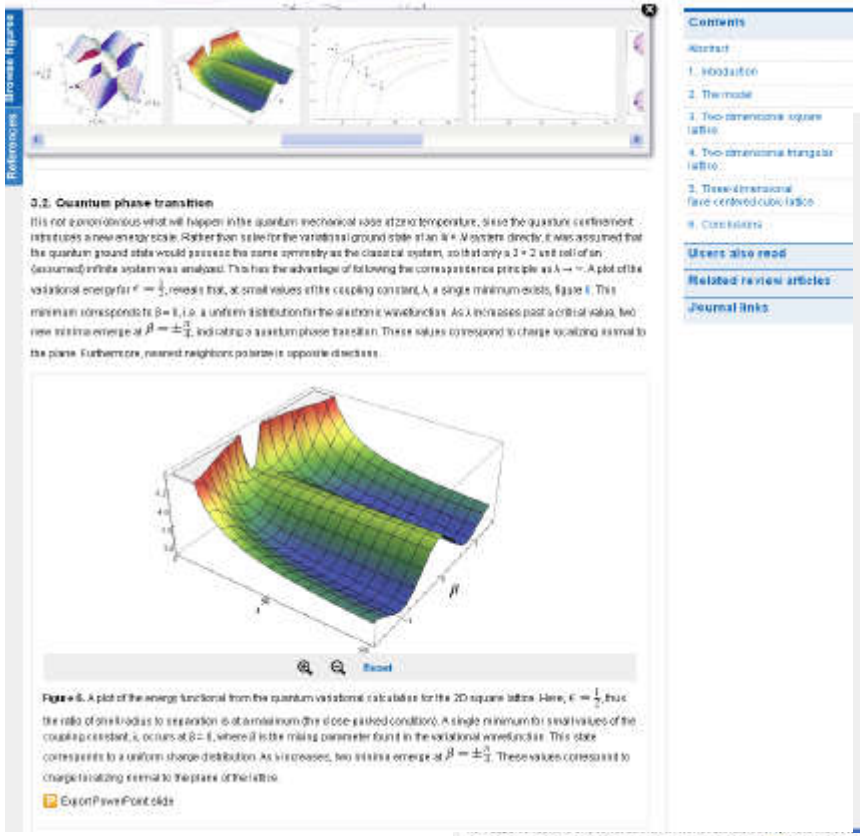
2. Semantic Publishing

- **Enriched Publications** **Data in a paper can be opened in Excel**
  - **Enhanced Publications: OECD Factbook**

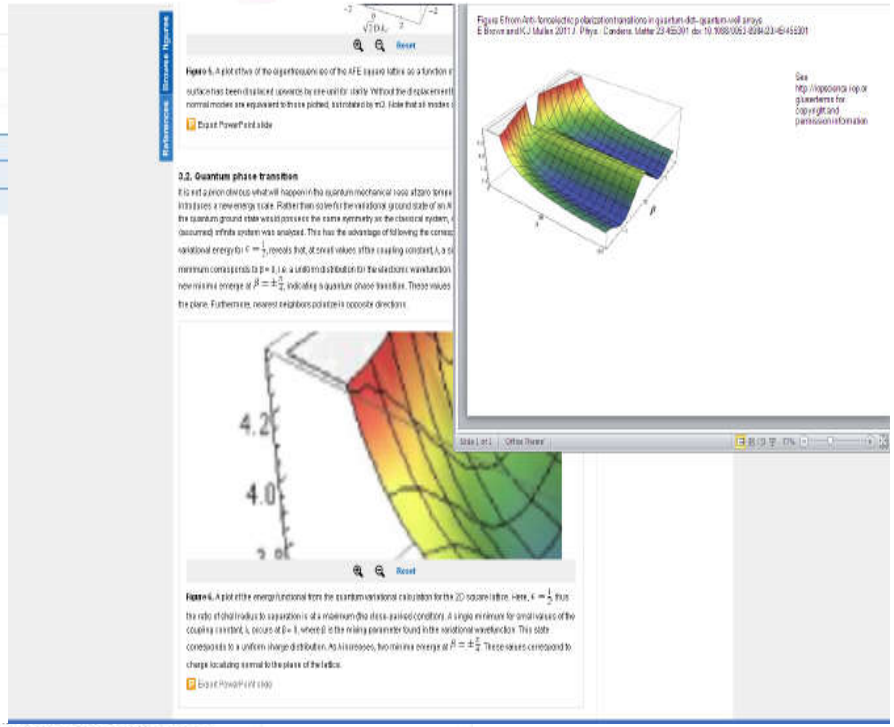


2. Semantic Publishing

Enriched Publications



Graphics zoom and .PPT



RIP - Windows Internet Explorer

http://www.cs.uu.nl/research/projects/i-cult/xposre/demo/gallows/

文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)

2 / 19 What is in a name | Table of Contents | PDF | Screen | Link to This | Source | Status | Format

Gallows in Medieval Frisia

J.A. Mol

URL http://www.cs.uu.nl/research/projects

version 1.1.1 - March 2011

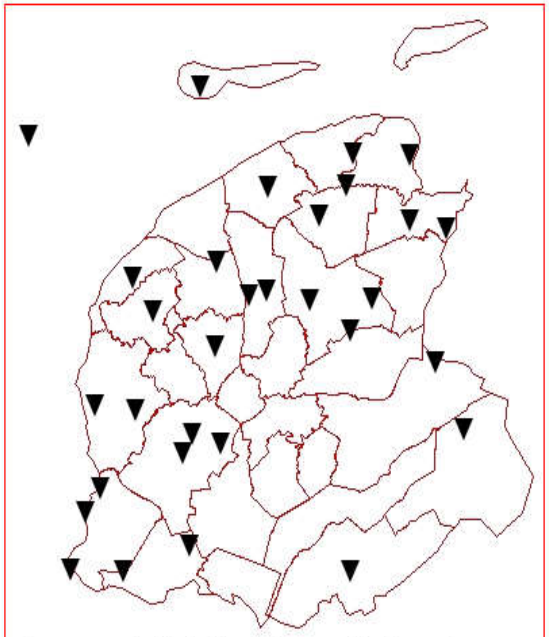
What is in a name

Gallows are often memorialised in the names of fields, water courses and other landscape features, such as in:

- Galge(n)berg (Gallows Hill)
- Galgeduin(Gallows Dune)
- Galgefenne (Gallows Fen)
- Galgekamp (Gallows Pitch)
- Galgerak (Gallows Reach)
- Galgeveld (Gallows Field)
- Galgewater (Gallows Water)
- Galgewier (Gallows Rise)
- Galgeweel (Gallows Pool).

These names have retained their gallows component after the eponymous construction had fallen into disuse and had subsequently decayed or been demolished. These names can be inventoried and localised, after which it can be concluded whether they indeed refer to the former presence of a gallows.

Such names represent the primary material for this essay. I have assembled them with the help of the huge collection of names of fields and water bodies [2] put together at the Frisian Academy in the 1940s and 1950s. Important in this connection is the institute's historical Geographic Information System (GIS), because it allows us not only to locate the physical position of gallows but also their owners, thereby giving some indication of who may have been involved in the setting up and maintenance of the gallows.

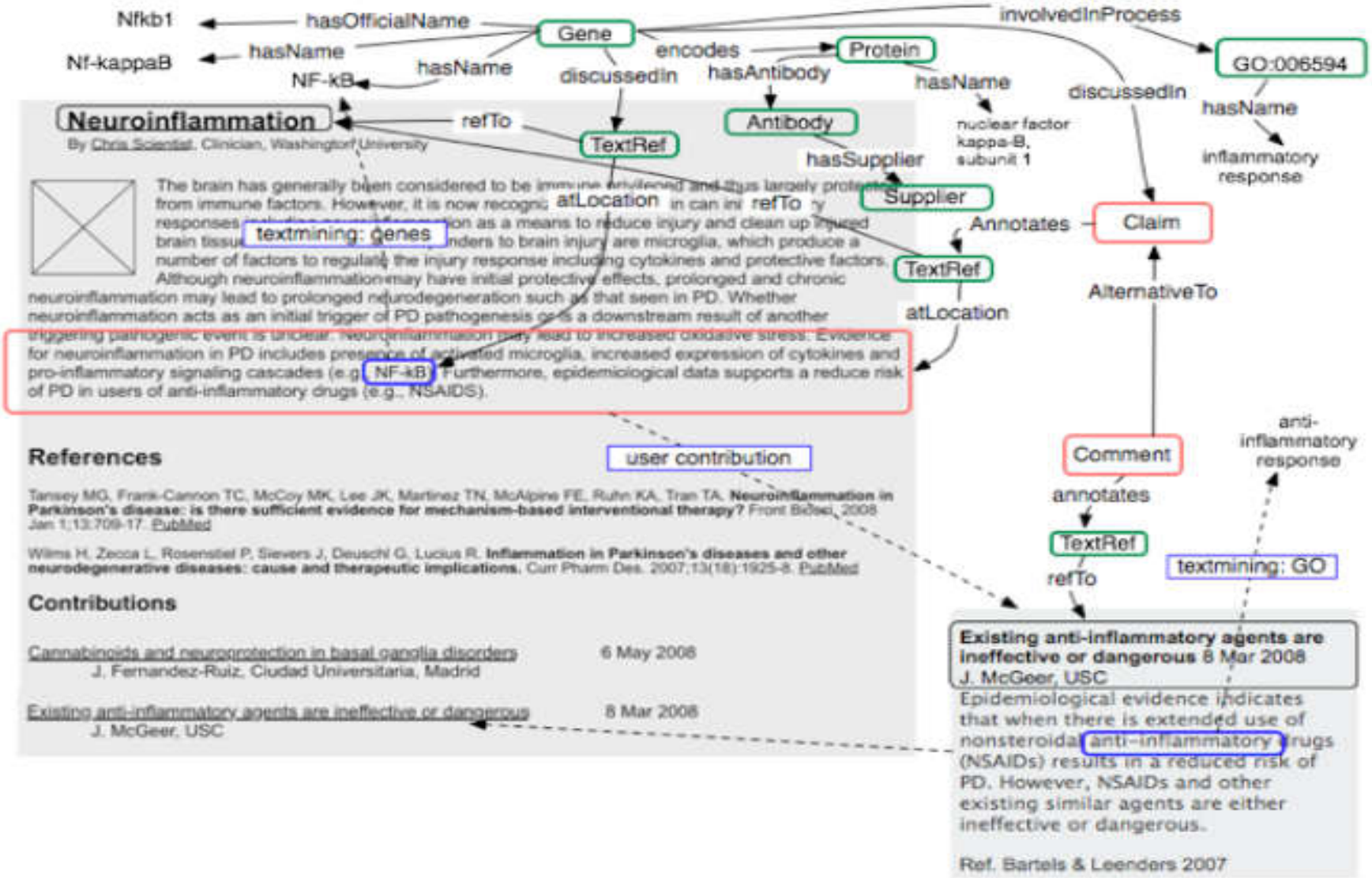


Gallows as recorded in the historical geographis information system HISGIS.

Try HISGIS yourself and find out how to locate gallows and other landmarks, or go to the tutorial on the GIS Pilot (Meertens Institute) to search for gallows using

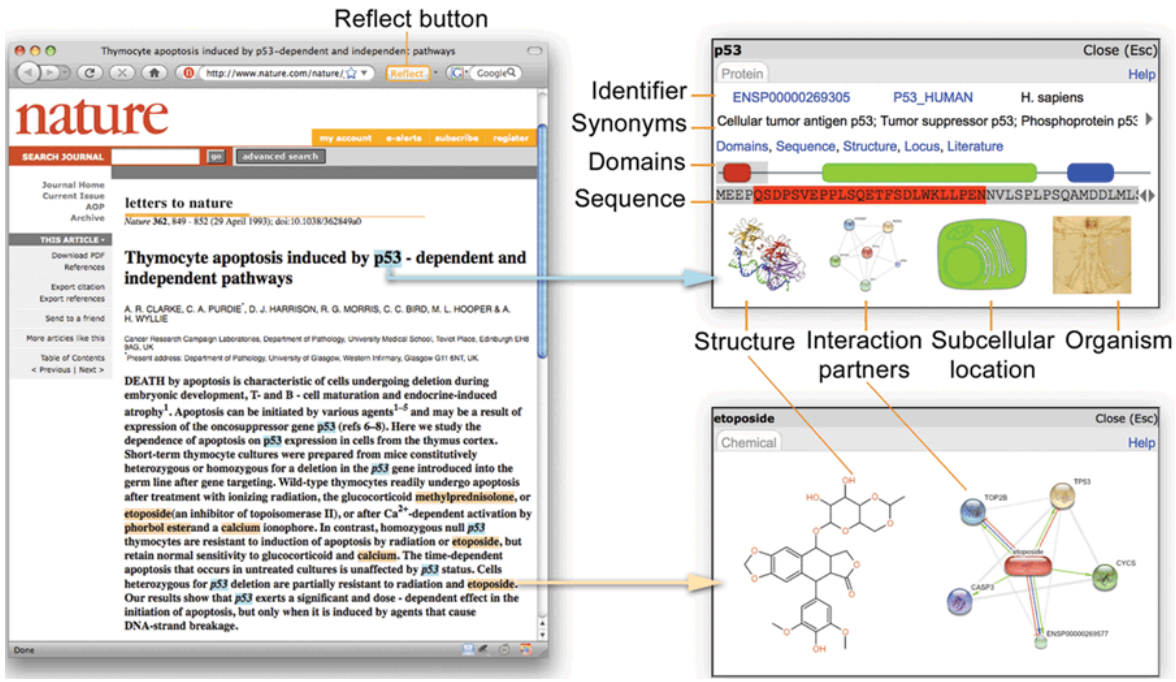
完成 Internet 100%

- **Semantic tagging of journal articles. Sudeshna Das, June 10, 2009**



2. Semantic Publishing

- Reflect
- creates a view of the web tailored for the life scientist, that is, with systematic tagging of biochemical entities, and easy access to more detailed information.
- Reflect is already being used by thousands of researchers.



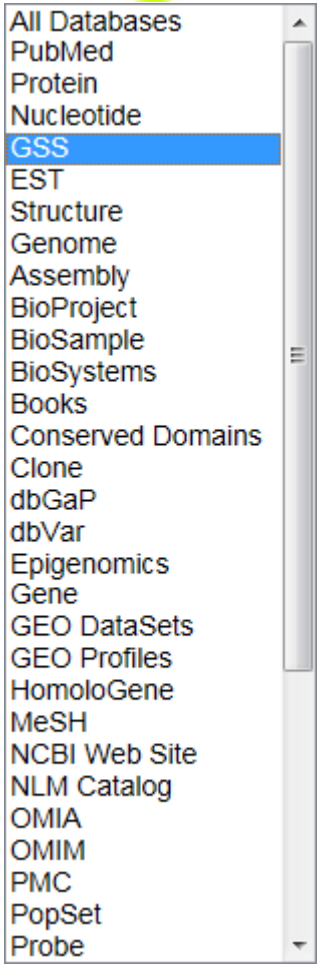
1. Pafilis E.，O'Donoghue S.I.，Jensen L.J.，Horn H.，Kuhn M.，Brown N.P. and Schneider R. Reflect — augmented browsing for the life scientist[J]. Nature Biotechnology，2009(27): 508-510

The 3rd form of interoperation

- Integrated services
  - *linking data and literature via integrated search and exploration services*

3. Integrated services

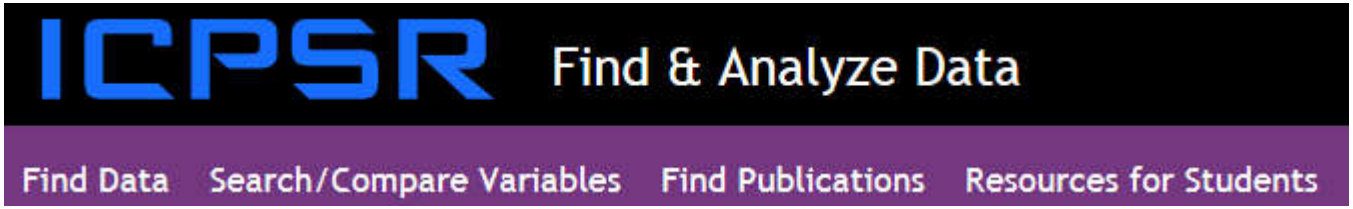
- Traditional web-based integrated search system
  - An example: NCBI Entrez Global Query
    - an integrated search and retrieval system that provides access to all databases simultaneously with a single query string and user interface
    - Entrez can efficiently retrieve related sequences, structures, and references.



1. http://www.ncbi.nlm.nih.gov/sites/entrez

3. Integrated services

- Traditional web-based integrated search system
  - *Similarily, ICPSR Bibliography of Data-related Literature*
  - a searchable database that contains over 60,000 citations of known published and unpublished works resulting from analyses of data held in the ICPSR archive.



1. http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/index.jsp

3. Integrated services

- Traditional web-based integrated search system
  - Another example: integrated search service in CAS
  - Integration of
    - searching scientific data in Chinese Scientific Database
    - searching literature resources (articles, patents, standards, and …) in National Science Library

Search databases

Search data in databases

Search papers

Search books

跨界集成检索

您现在的位置：>首页>>数据服务>>查找数据>>跨界集成检索

>>跨界集成检索

中国科学院科学数据库： 数据库 数据库中数据

中国科学院国家科学图书馆： 图书 标准 学位论文 维普期刊论文 网络科技期刊论文

野外台站数据库： 中国陆地生态系统碳氮水循环综合数据库数据集

任意字段 二次检索 重新检索

检索斑头雁获得元数据1条记录，以下是1-1条记录

数据库名称：	青海湖鸟类GPS跟踪数据库	详细信息 访问此库 咨询
数据库简介：	目前青海湖候鸟的跟踪同时使用GPS(全球卫星定位系统)和多普勒平移两种定位装置。青海湖候鸟跟踪的总数为77只，其中斑头雁29只，携带了45g的加强型PTT，其中包含了GPS接收机，可以同时通过Argos卫星和GPS接收机进行定位；同样采用两种方法定位的是16只赤麻鸭，携带的是30g的PTT。另外还有鱼鸥10只，针尾鸭1只和棕头鸥1只携带的是18g 的Northstar PTT，只通过Argos卫星进行定位。目前，鸟类的监测卫星数据，先由地面接收站将接收到的Argos卫星的DIAG数据进行初步处理发送到美国地质勘探局，经过研究人员的一次处理之后，再将数据发送给中国科学院计算机网络信息中心。结合Argos卫星的数据标准以及地理信息学中的相关知识，我们对数据进行了进一步的处理，并把数据存入关系数据库中。整理过后的数据记录主要包括以下几个主要字段：Animal字段表示被跟踪候鸟的唯一编号，Record_id表示数据获取的类型，用于区分不同跟踪设备的不同的获取地理信息的方法，Datetime字段表示所获取到数据的时间，Latitude和Longitude字段表示经度和纬度，Lc94字段用来标记数据的卫星位置等级，对于使用GPS进行定位的数据的级别为LG，使用Argos系统进行定位的数据等级分为7个级别，按照准确度增加的顺序分别是：Z、B、A、0、1、2和3，标记为LZ、LB、LA、L0、L1、L2和L3。按照候鸟迁徙研究中对精度的要求，在我们的数据挖掘过程中使用数据为LG、L1～L3这四种精度的数据。	

3. Integrated services

- New opportunity: Linking data and literature in Linked Data context
  - Linked Data: Tim Berners-Lee coined the term Linked Data in 2006[1].
  - Based on the concept of linked data, W3C initiated the Linking Open Data movement.
  - It has driven many data sets which are distributed in more than 200 domains published as Linked Data.

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

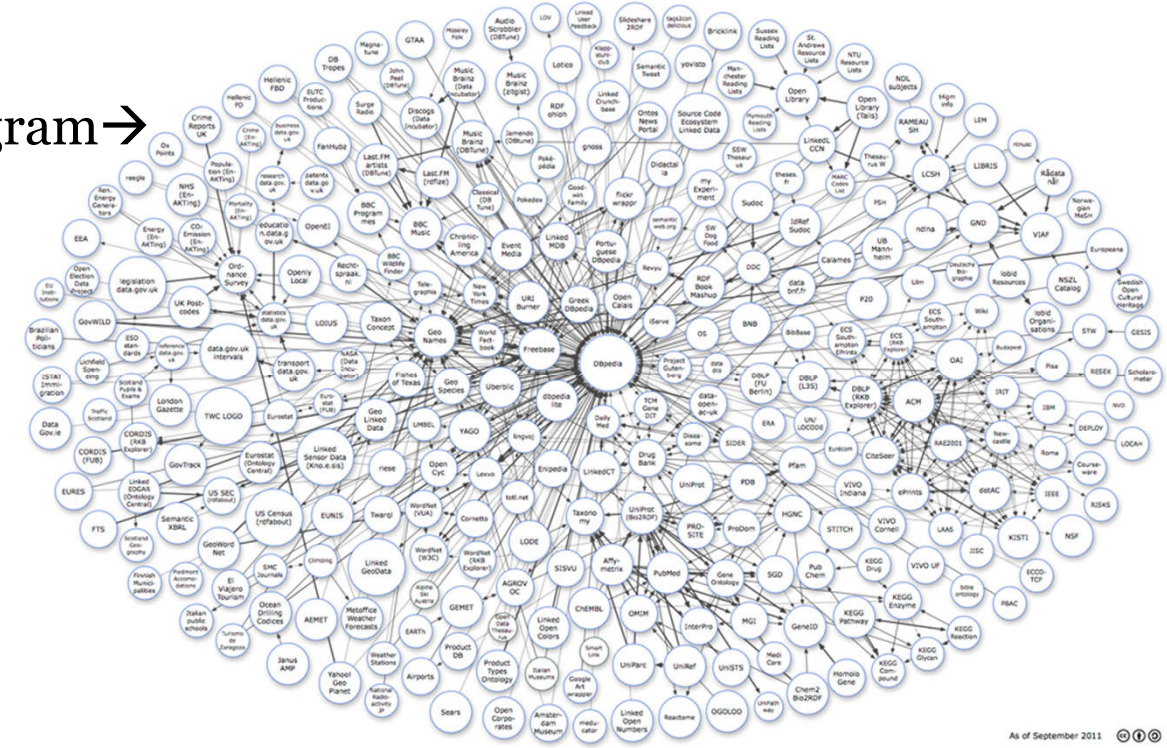
3. Integrated services

- 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.

3. Integrated services

- By September 2011, LOD had covered about 31 billion RDF triples and about 500 million RDF links.

LOD dataset cloud diagram→



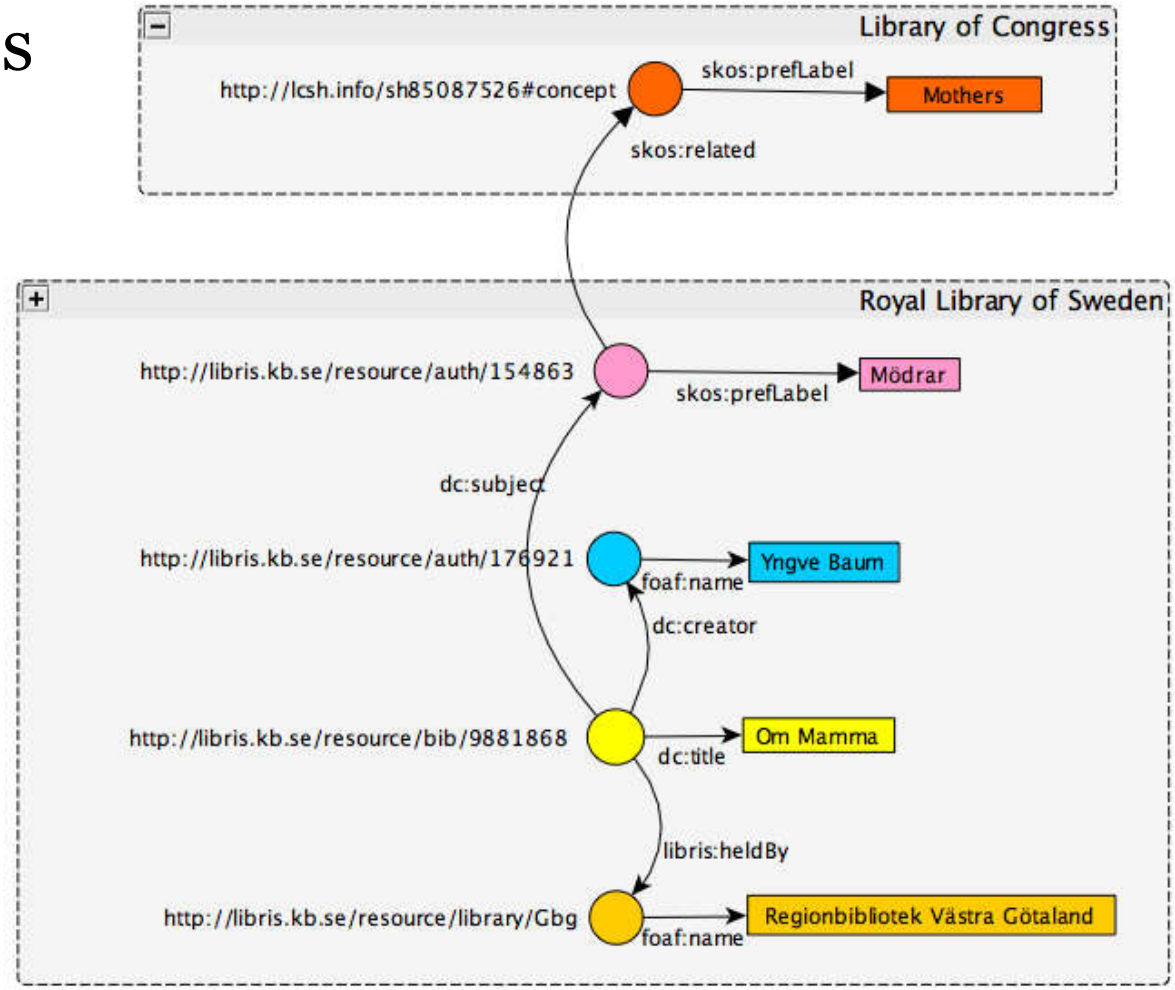
1. Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

3. Integrated services

- Publishing literature resources as Linked Data
  - Classification systems
  - Subject headings/subject authority files
  - Name authority data
  - Thesauri
  - Other controlled vocabularies
- Libraries
  - Royal Library of Sweden: LIBRIS
  - Library of Congress: LCSH
  - German National Library
  - National Library of France（BnF）
  - Hungarian National Library

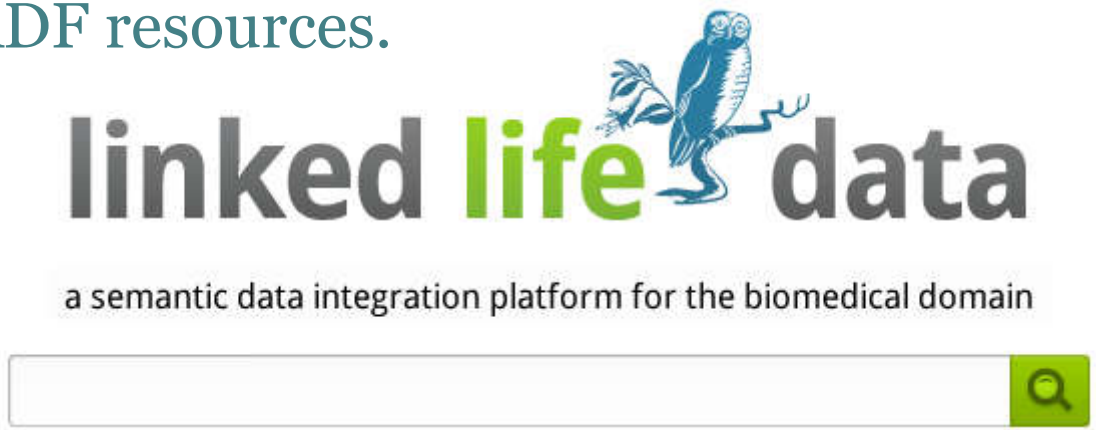
- Library resources are linked through RDF links

- dc:subject of an article refers to a skos:Concept defined in LCSH (Library of Congress Subject Headings)



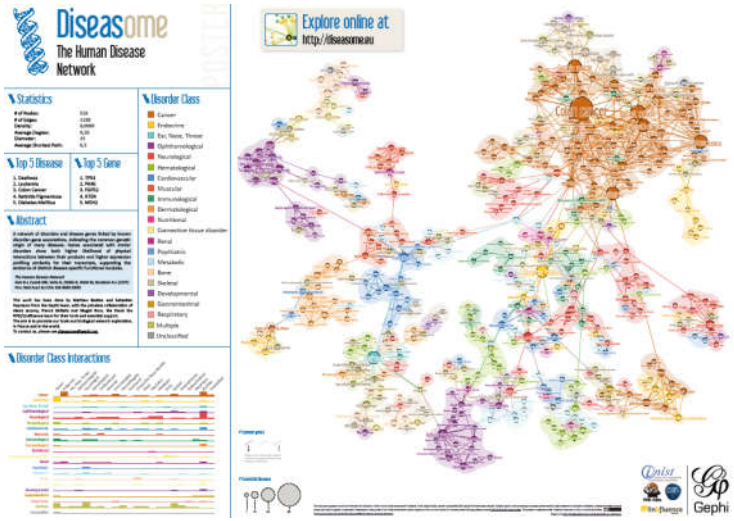
3. Integrated services

- Publishing Scientific Data as Linked Data
  - Search and explore over RDF statements from various sources including UniProt, PubMed, EntrezGene and 20 more…
  - Perform complex SPARQL queries and retrieve 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.

3. Integrated services



- 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
  - is 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.

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

3. Integrated services

- GeoSpecies Knowledge Base
  - Publishing information on Biological Orders, Families, Species as well as species occurrence records and relate data, links to geonames, bio2rdf, dbpedia, freebase, umbel.

GeoSpecies Knowledge Base

Kingdom	Common Name
Animalia	Animals
Plantae	Plants
Fungi	Fungi
Protozoa	Protozoans
Bacteria	Bacteria
Archaea	Archaea
Chromista	Chromista
Viruses	Viruses

1. http://lod.geospecies.org/

3. Integrated services

- Linked Data make it possible to link scientific data and literature resources as same format

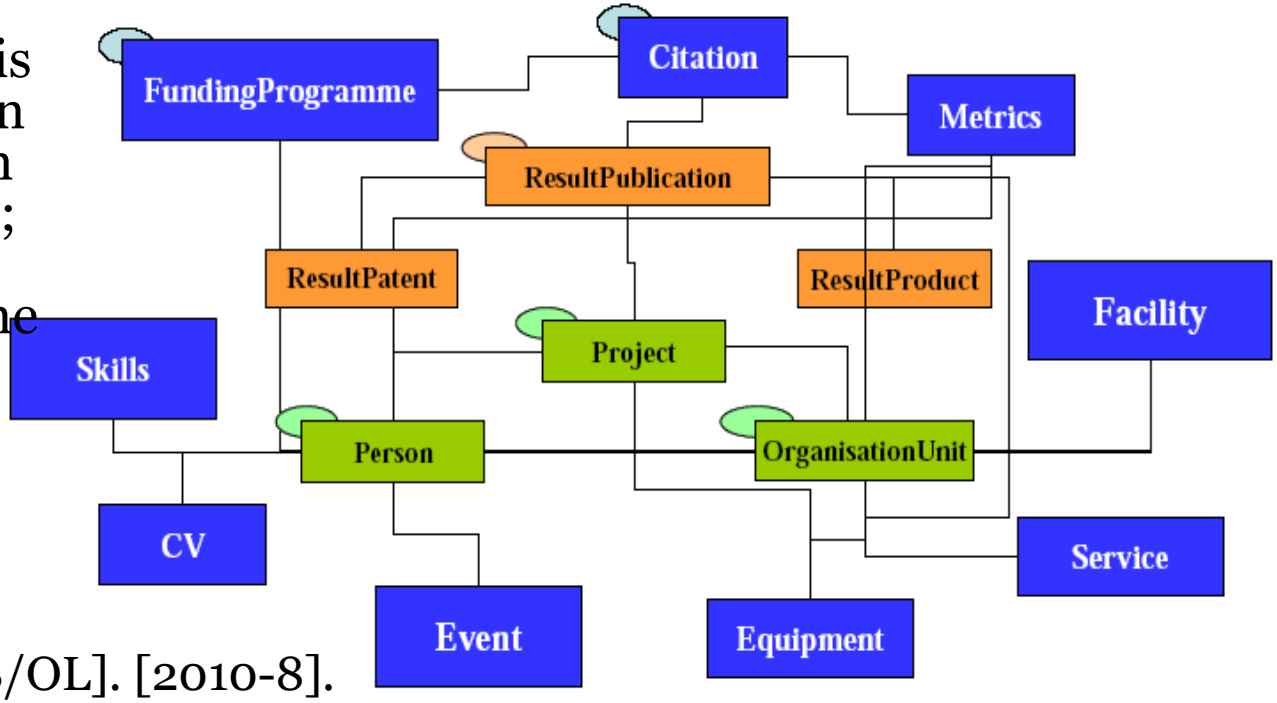
- Therefore researchers focused on building semantic data models for representation the relationship of scientific data, literature and more RESEARCH OBJECTs

1. http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

euroCRIS
Current Research Information Systems

3. Integrated services

- CERIF is a standard for managing and exchanging research data, ie information about researchers, organizations, projects, outputs and funding, arising from the research process. It provides a data model that can be used to describe the research domain, including relationships between the constituent parts, and how these change over time.
- Officially CERIF is a European Union Recommendation to member states; it was originally developed with the support of the European Commission.



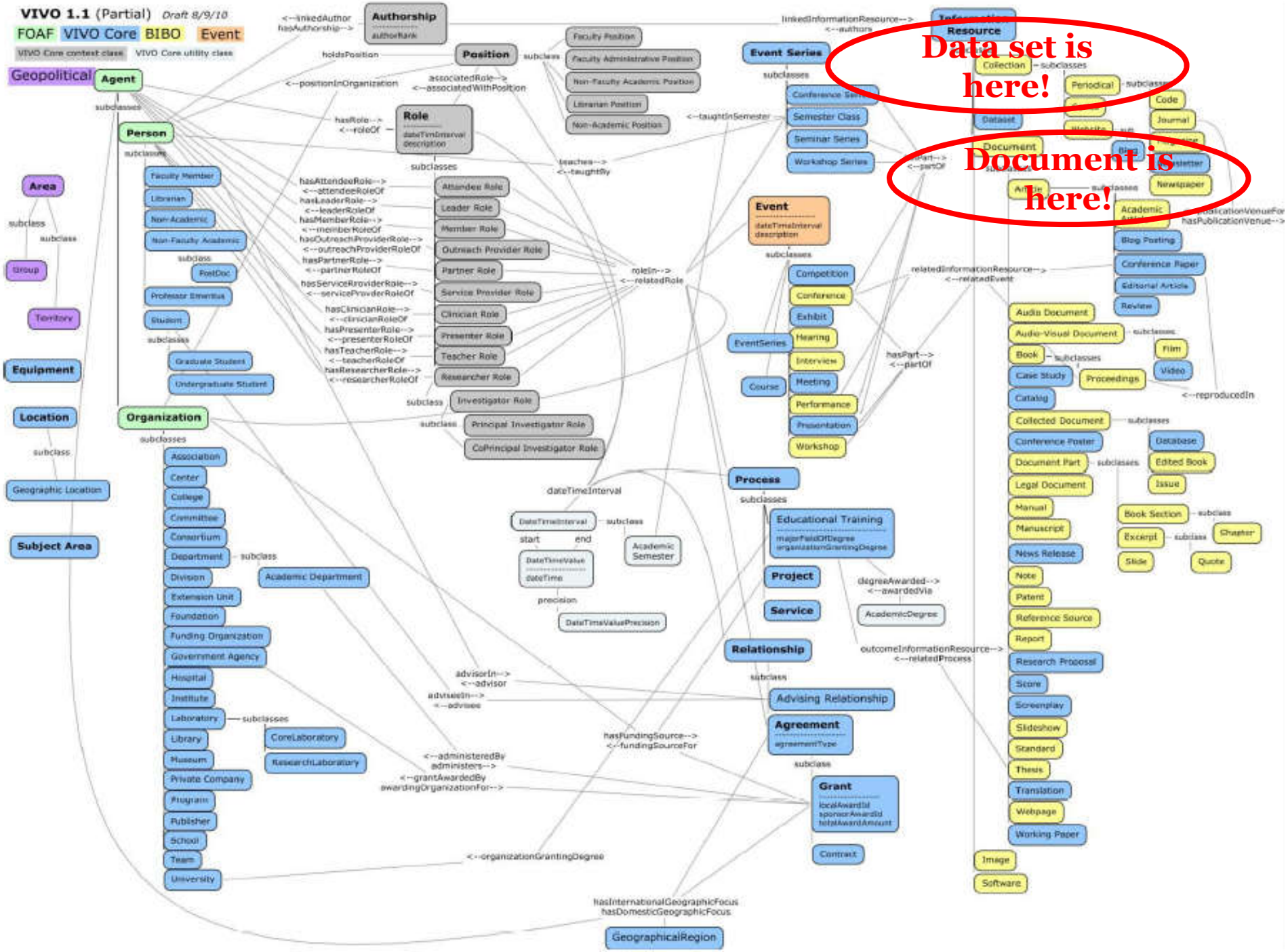
1. CERIF Releases. [EB/OL]. [2010-8]. http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

3. Integrated services

- Linking literature and scientific data in LOD context
  - VIVO
  - a tool for representing information about research and researchers -- their scholarly works, research interests, and organizational relationships.
  - VIVO provides an expressive ontology, tools for managing the ontology, and a platform for using the ontology to create and manage linked open data for scholarship and discovery.
  - By the end of 2012, over 20 countries and 50 organizations will provide information in VIVO format on more than one million researchers and research staff, including publications, research resources, events, funding, courses taught, and other scholarly activity.

VIVO Research & Expertise Across Cornell

1. http://vivo.cornell.edu


VIVO 1.1 (Partial) Draft 8/9/10
FOAF VIVO Core BIBO Event
VIVO Core context class
VIVO Core utility class
Geopolitical
Agent
Person
Area
Group
Territory
Equipment
Location
Geographic Location
Subject Area
Organization
Authorship
Position
Role
Faculty Position
Faculty Administrative Position
Non-Faculty Academic Position
Librarian Position
Non-Academic Position
Faculty Member
Librarian
Non-Academic
Non-Faculty Academic
PostDoc
Professor Emeritus
Student
Graduate Student
Undergraduate Student
Attendee Role
Leader Role
Member Role
Outreach Provider Role
Partner Role
Service Provider Role
Clinician Role
Presenter Role
Teacher Role
Researcher Role
Investigator Role
Principal Investigator Role
CoPrincipal Investigator Role
Association
Center
College
Committee
Consortium
Department
Academic Department
Division
Extension Unit
Foundation
Funding Organization
Government Agency
Hospital
Institute
Laboratory
CoreLaboratory
ResearchLaboratory
Library
Museum
Private Company
Program
Publisher
School
Team
University
Event Series
Conference Series
Semester Class
Seminar Series
Workshop Series
Event
Competition
Conference
Exhibit
Hearing
Interview
Meeting
Performance
Presentation
Workshop
Course
Process
Educational Training
Project
Service
Relationship
Advising Relationship
Agreement
Grant
Contract
Resource
Collection
Dataset
Document
Article
Periodical
Code
Journal
Website
Blog
Newsletter
Newspaper
Academic Article
Blog Posting
Conference Paper
Editorial Article
Review
Audio Document
Audio-Visual Document
Film
Video
Book
Case Study
Proceedings
Catalog
Collected Document
Database
Conference Poster
Document Part
Edited Book
Legal Document
Issue
Manual
Book Section
Manuscript
Excerpt
Chapter
News Release
Slide
Quote
Note
Patent
Reference Source
Report
Research Proposal
Score
Screenplay
Slideshow
Standard
Thesis
Translation
Webpage
Working Paper
Image
Software
DateTimeInterval
DateTimeValue
DateTimeValuePrecision
Academic Semester
AcademicDegree
GeographicalRegion
Data set is here!
Document is here!

Summary

- This paper reviews interoperation between scientific data and literature in three forms:
  1. Data publication and citation
     publishing and citing scientific data like papers
     - *Forms*
     - *Publication Framework*
     - *Organizations invovled*
  2. Semantic Publishing
     publishing actionable data in articles
     - *Concept*
     - *Forms*
     - *Applications and tools*
  3. Integrated services
     linking data and literature via integrated search and exploration services
     - *Traditional integrated search and exploration system*
     - *Linking data and literature in LOD context*

Thank you!