Outline of the Tutorial Modules

I. Overview on Scientific Data Management (Gertz)
13:30—14:15

II. From Conventional to Scientific Data Integration (Ludaescher)
14:15—15:00

III. From Scientific Data Formats to Data Stream Processing (Gertz)
15:30—16:15

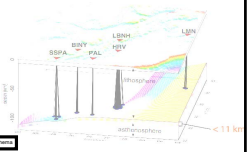
IV. Introduction to Scientific Workflows (Ludaescher)
16:15-17:00

II. From Conventional to Scientific Data Integration

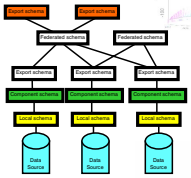
- Overview
  - From Science to e-Science to Cyberinfrastructure
  - Conventional (schema-based) data integration
  - Scientific DI using ontology-based extensions
  - From metadata to ontologies
  - Link-based 'integration'

Types of "Integration"

- Spatial (co-)registration/"overlay" of different data
  - from 2D, 3D, 4D (x,y,z,*t*), (4+n) D

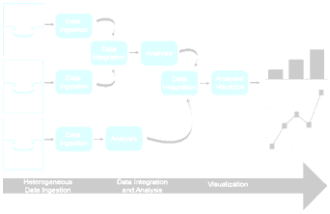


- **Conventional (DB-oriented) integration:**
  - schema-based
  - view-based
  - at the data-level



- **Extended DI approaches using "ontologies"**
  - ontologies? controlled vocabularies? metadata/annotations?

- Application/process integration
  - ➔ scientific workflows (Module 4)



- Other mechanisms of "integration"
  - link-based, clustering, …
  - statistics, data mining, visualization, …

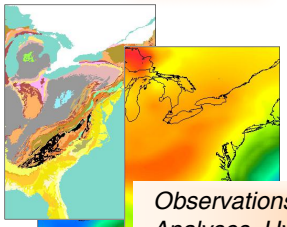
Different Science Domains

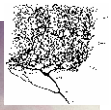
Natural Sciences

Earth Sciences

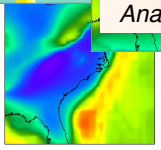
Physical Sciences

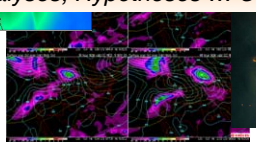
Life Sciences

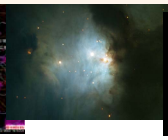


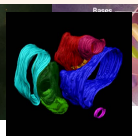


Observations, Measurements, Models, Simulations, Analyses, Hypotheses … Understanding, Prediction







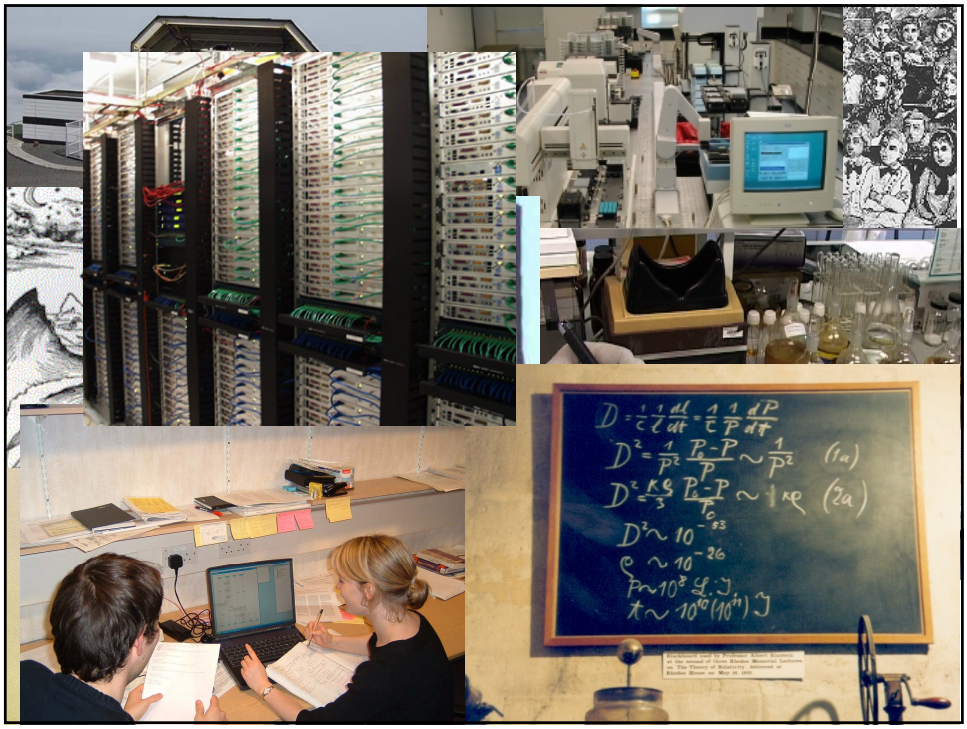


data-intensive

compute-intensive

metadata-intensive

structurally & semantics -intensive



e-Science (UK) and Cyberinfrastructure (US)

- *"**e-Science** is about global **collaboration** in key areas of **science** and the next generation of **[computing] infrastructure** that will enable it."*

  Sir John Taylor, Director Office of Science and Technology, UK

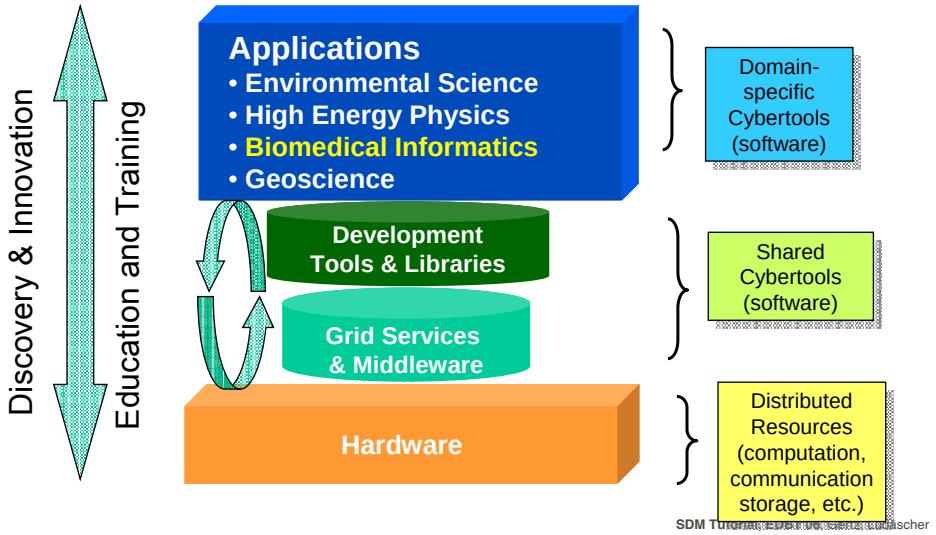
- *"**Cyberinfrastructure** is the coordinated aggregate of software, hardware and other technologies, as well as human expertise, required to support current and future discoveries in science and engineering. The challenge of Cyberinfrastructure is to **integrate** relevant and often disparate resources to provide a useful, usable, and enabling framework for research and discovery characterized by broad access and 'end-to-end' coordination."*

  Fran Berman, San Diego Supercomputer Center, UCSD

Integrated Cyberinfrastructure System

meeting the needs of multiple communities

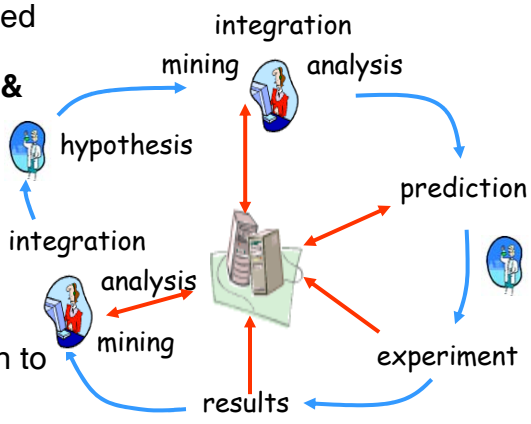
Source: Dr. Deborah Crawford, Chair, NSF CyberInfrastructure Working Group



SDM Tutorial, [illegible] Ludäscher

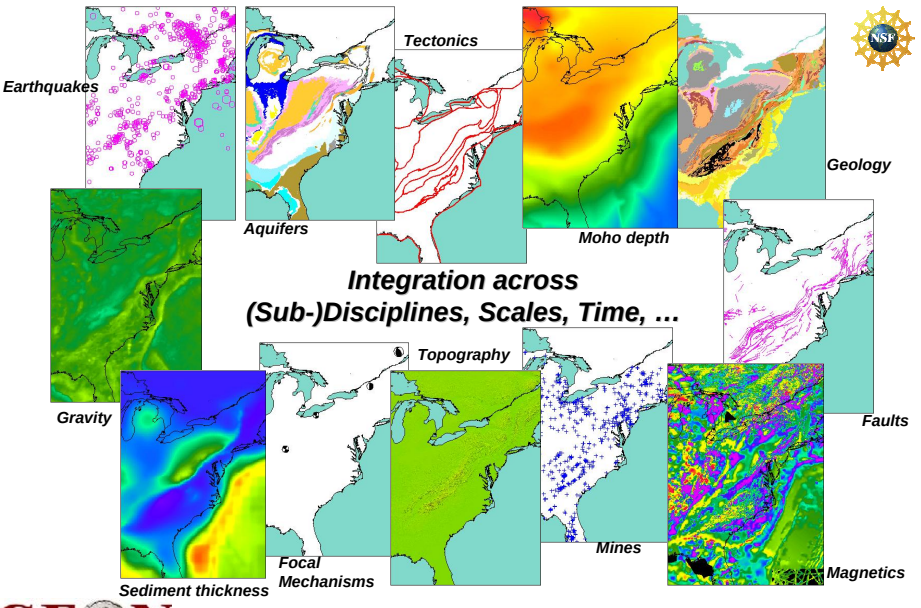
Science + e-Science

- Discovery increasingly done **in silico** on results obtained from experiments using **computational analysis & data repositories**.
- A new era of **collection based** and **simulation based** science, in addition to hypothesis driven and experimental science



Src: Carole Goble ISWC'05 Keynote

So much data, so little time …

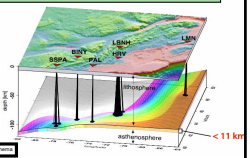


GEON CYBERINFRASTRUCTURE FOR THE GEOSCIENCES

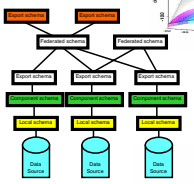
Src: Dogan Seber SDSC, Krishna Sinha VT

Types of "Integration"

- **Spatial (co-)registration/"overlay" of different data**
  - from 2D, 3D, 4D (x,y,z,*t*), (4+n) D

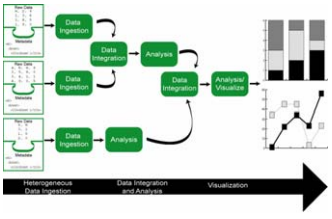


- **Conventional (DB-oriented) integration:**
  - schema-based
  - view-based
  - at the data-level



- **Extended DI approaches using "ontologies"**
  - ontologies? controlled vocabularies? metadata/annotations?

- **Application/process integration**
  - ➔ scientific workflows (Module IV)

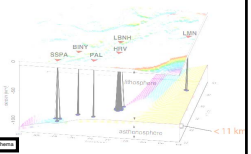


- **Other mechanisms of "integration"**
  - link-based (Aladin), clustering, …
  - statistics, data mining, visualization, …

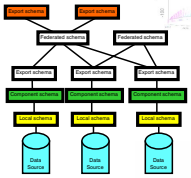
SDM Tutorial, EDBT'06, Gertz, Ludäscher

Types of "Integration"

- Spatial (co-)registration/"overlay" of different data
  - from 2D, 3D, 4D (x,y,z,*t*), (4+n) D

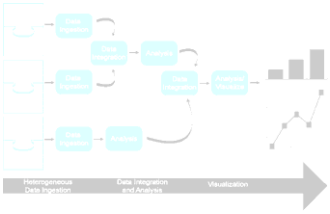


- **Conventional (DB-oriented) integration:**
  - schema-based
  - view-based
  - at the data-level



- **Extended DI approaches using "ontologies"**
  - ontologies? controlled vocabularies? metadata/annotations?

- Application/process integration
  - ➔ scientific workflows (Module 4)



- Other mechanisms of "integration"
  - link-based, clustering, …
  - statistics, data mining, visualization, …

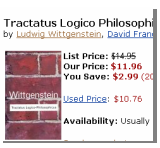
An Online Shopper's Information Integration Problem

El Cheapo: "Where can I get the cheapest copy (including shipping cost) of Wittgenstein's Tractatus Logicus-Philosophicus within a week?"

addall.com

Mediator (virtual DB)
(vs. Datawarehouse)
NOTE: non-trivial data engineering challenges!

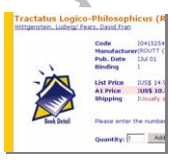
"One-World" Mediation



amazon.com

barnes&noble.com

half.com



A1books.com

A Home Buyer's Information Integration Problem

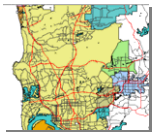
What houses for sale under $500k have at least 2 bathrooms, 2 bedrooms, a nearby school ranking in the upper third, in a neighborhood with below-average crime rate and diverse population?

?

Information Integration

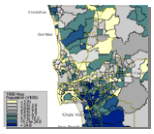
"Multiple-Worlds" Mediation

Realtor



Crime Stats

School Rankings



Demographics

A Neuroscientist's Information Integration Problem

Biomedical Informatics Research Network
http://nbirn.net

What is the cerebellar distribution of rat proteins with more than 70% homology with human NCS-1? Any structure specificity? How about other rodents?

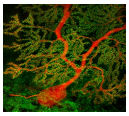
?

Information Integration

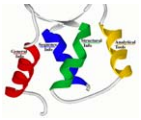
"Complex Multiple-Worlds" Mediation

Inter-source links:

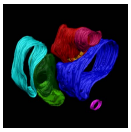
- *unknown to the non-scientists*
- *unclear for the scientist*



protein localization (NCMIR)



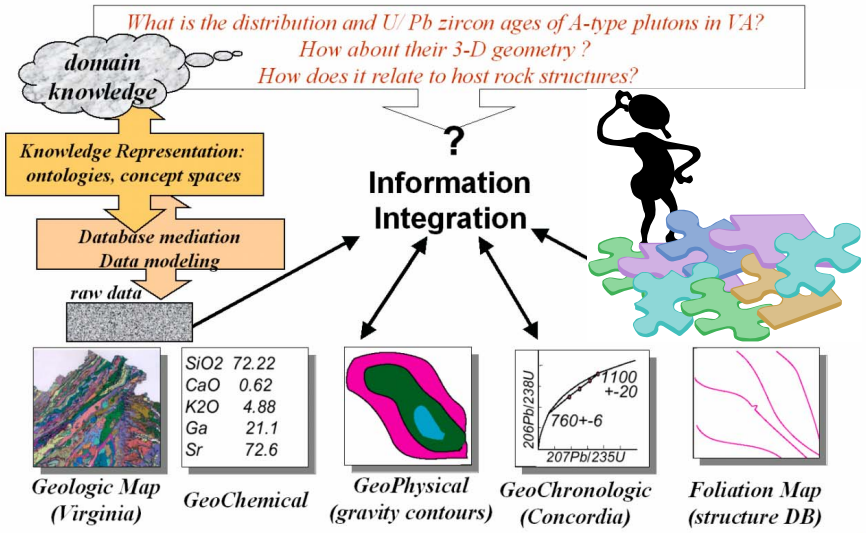
sequence info (CaPROT)



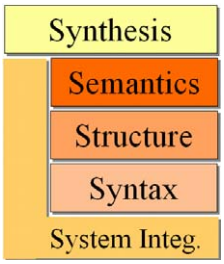
morphometry (SYNAPSE)

neurotransmission (SENSELAB)

The Problem: Scientific Data Integration
or: … *from Questions to Queries* …



Interoperability & Integration Challenges

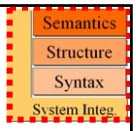


- reconciling S^5 heterogeneities
- "gluing" together resources
- bridging information and knowledge gaps computationally

- ***System aspects: "Grid" Middleware***
  - distributed data & computing, SOA
  - resource discovery, authentication, authorization
  - web services, WSDL/SOAP, WSRF, OGSA, …
  - *(re-)sources = services, files, data sets, nodes …*
- ***Syntax & Structure:***
  (XML-Based) Data Mediators
  - wrapping, restructuring
  - (XML) queries and views
  - *sources = (XML) databases*
- ***Semantics:***
  Model-Based/Semantic Mediators
  - conceptual models and declarative views
  - Knowledge Representation: ontologies, description logics (RDF(S),OWL ...)
  - *sources = knowledge bases (DB+CMs+ICs)*
- ***Synthesis: Scientific Workflow Design & Execution***
  - Composition of declarative and procedural components into larger workflows
  - *(re)sources = services, processes, actors, …*

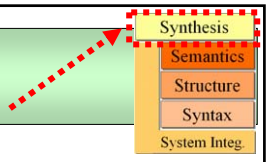
SDM Tutorial, EDBT'06, Gertz, Ludäscher

Information Integration Challenges: *S⁴ Heterogeneities*



- ***S*ystem aspects**
  - platforms, devices, data & service distribution, APIs, protocols, …
  - ➔ **Grid middleware technologies**
  - + e.g. single sign-on, platform independence, transparent use of remote resources, …
- ***S*yntax & *S*tructure**
  - heterogeneous data formats (*one for each tool* ...)
  - heterogeneous data models (*RDBs, ORDBs, OODBs, XMLDBs, flat files, …*)
  - heterogeneous schemas (*one for each DB* ...)
  - ➔ **Database mediation technologies**
  - + XML-based data exchange, integrated views, transparent query rewriting, …
- ***S*emantics**
  - descriptive metadata, different terminologies, "hidden" semantics (context), implicit assumptions, …
  - ➔ **Knowledge representation & semantic mediation technologies**
  - + "smart" data discovery & integration
  - + e.g. ask about ***X*** ('**mafic**'); find data about ***Y*** ('**diorite**'); be happy anyways!

Information Integration Challenges: *S⁵ Heterogeneities*



- ***S*ynthesis** of applications, analysis tools, data & query components, … into "**scientific workflows**"
  - How to put together components to solve a scientist's problem?

➔ **Scientific Problem Solving Environments (PSEs)**

➔ **Portals, Workbench ("scientist's view")**

+ ontology-enhanced data registration, discovery, manipulation

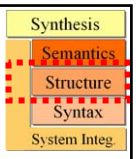
+ creation and registration of new data products from existing ones, …

➔ **Scientific Workflow System ("engineer's view")**

+ for designing, re-engineering, deploying analysis pipelines and scientific workflows; ***a tool to make new tools …***

+ e.g., creation of new datasets from existing ones, dataset registration, …

Information Integration from a DB Perspective



- **Information Integration Problem**
  - **Given:** data sources S_1, ..., S_k (databases, web sites, ...) and user questions Q_1,..., Q_n that can –in principle– be answered using the information in the S_i
  - **Find:** the answers to Q_1, ..., Q_n
- **The Database Perspective:** source = "database"

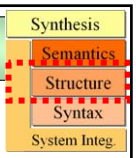
  ⇒ S_i has a **schema** (relational, XML, OO, ...)

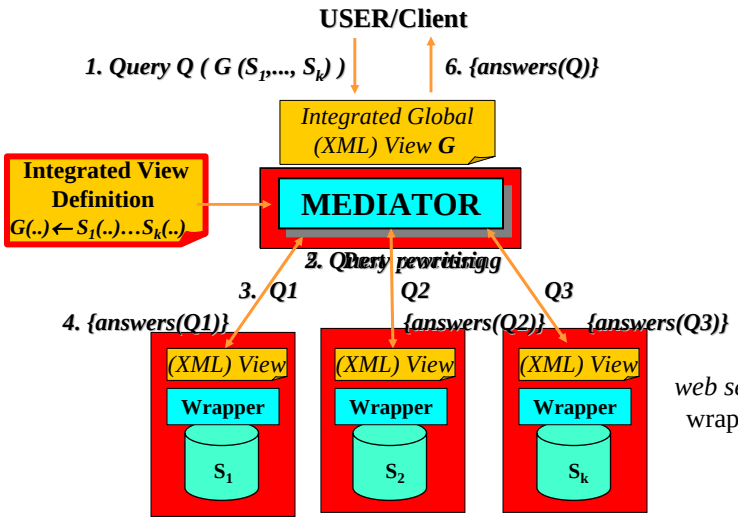
  ⇒ S_i **can be queried**

  ⇒ define virtual (or materialized) **integrated (or global) view G** over local sources S_1 ,..., S_k using **database query languages** (SQL, XQuery,...)

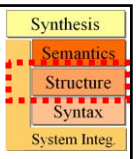
  ⇒ ***questions become queries*** Q_i against G(S_1,..., S_k)

Standard (XML-Based) Mediator Architecture





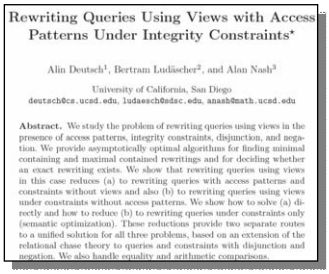
Query Rewriting in Data Integration



- **Given**:
  - **Declarative user query Q**: ***answer(…) :– …G …***
  - … & { G **:–** … L … } *global-as-view* (GAV)
  - … & { L **:–** … G … } *local-as-view* (LAV)
  - … & { ic_false(…) **:–** … L … G… } *integrity constraints* (ICs)
- **Find**:
  - equivalent (or *minimal containing*, *maximal contained*) **query plan Q'**: ***answer(…) :– … L …***
  - ➔ query rewriting
- **Results**:
  - A variety of results/algorithms; depending on classes of queries, views, and ICs: **P**, **NP**, … , **undecidable**
  - still a hot research area in the database community

Data Integration: Limited Access Patterns+Views+ICs

We study the problem of rewriting a query Q in terms of a given set of views $\mathcal{V}$ with limited access patterns $\mathcal{P}$, under a set Σ of integrity constraints. More precisely, we are interested in determining whether there exists a query plan Q', expressed in terms of the views $\mathcal{V}$ only, that is executable (i.e., observes $\mathcal{P}$) and equivalent to Q for all databases satisfying Σ. We say that Q is *feasible* if such Q' exists. For infeasible Q we seek the minima… contained executable queries, which provide the "best … plans for approximating the answer to Q from abo…



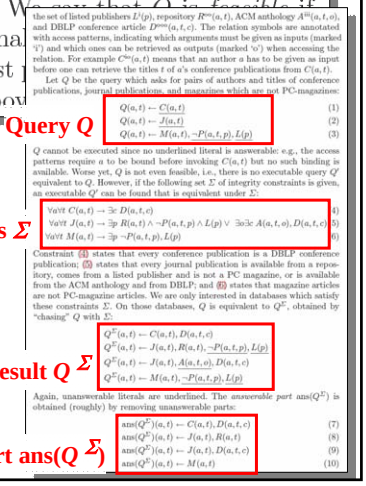
Rewriting Queries Using Views with Access Patterns Under Integrity Constraints*

Alin Deutsch[1], Bertram Ludäscher[2], and Alan Nash[3]

University of California, San Diego
deutsch@cs.ucsd.edu, ludaesch@sdsc.edu, anash@math.ucsd.edu

Abstract. We study the problem of rewriting queries using views in the presence of access patterns, integrity constraints, disjunction, and negation. We provide asymptotically optimal algorithms for finding minimal containing and maximal contained rewritings and for deciding whether an exact rewriting exists. We show that rewriting queries using views in this case reduces (a) to rewriting queries with access patterns and constraints without views and also (b) to rewriting queries using views under constraints without access patterns. We show how to solve (a) directly and how to reduce (b) to rewriting queries under constraints only (semantic optimization). These reductions provide two separate routes to a unified solution for all three problems, based on an extension of the relational chase theory to queries and constraints with disjunction and negation. We also handle equality and arithmetic comparisons.

ICDT'05



Query Q

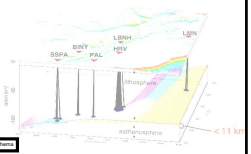
ICs Σ

Chase result Q^{Σ}

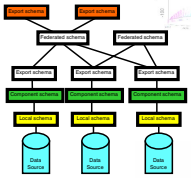
Answerable part ans(Q^{Σ})

Types of "Integration"

- Spatial (co-)registration/"overlay" of different data
  - from 2D, 3D, 4D (x,y,z,*t*), (4+n) D

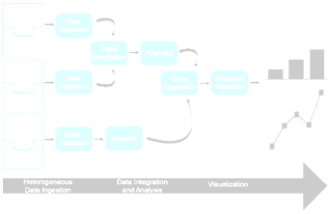


- **Conventional (DB-oriented) integration:**
  - schema-based
  - view-based
  - at the data-level



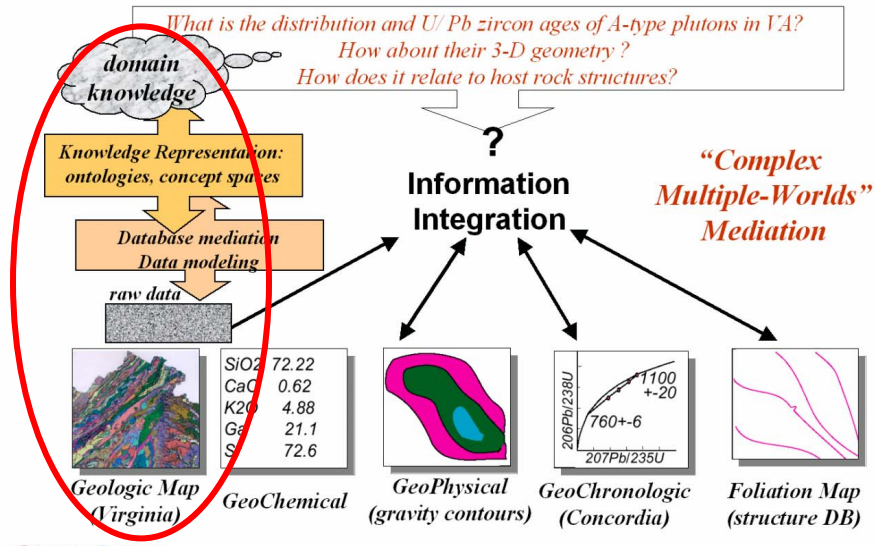
- **Extended DI approaches using "ontologies"**
  - ontologies? controlled vocabularies? metadata/annotations?

- Application/process integration
  - ➔ scientific workflows (Module 4)



- Other mechanisms of "integration"
  - link-based, clustering, …
  - statistics, data mining, visualization, …

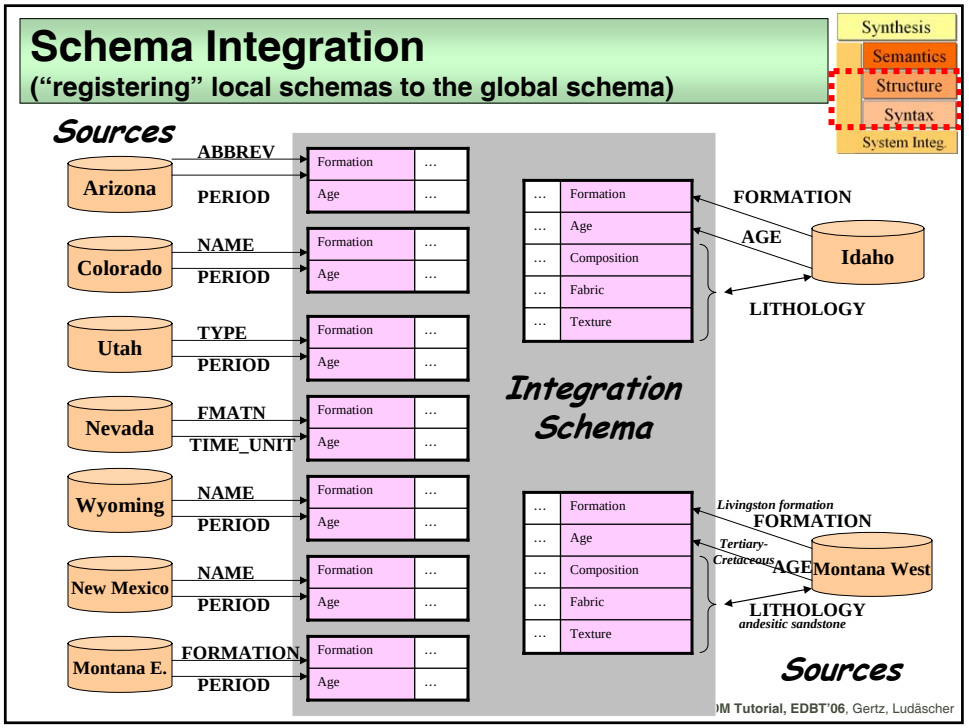
The Problem: Scientific Data Integration
or: *… from Questions to Queries …*



Example: Geologic Map Integration

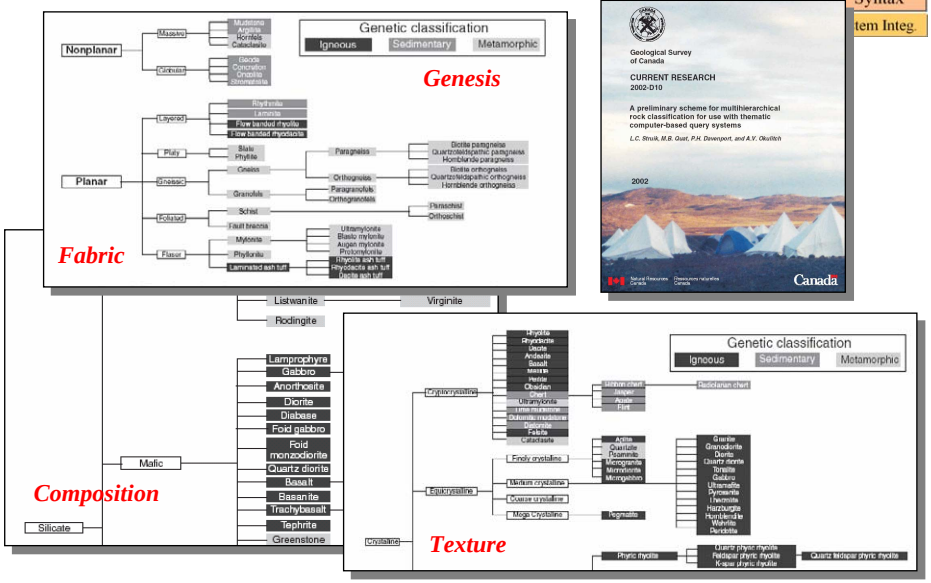
- **Given**:
  - Geologic maps from different state geological surveys (shapefiles w/ different data schemas)
  - Different ontologies:
    - Geologic age ontology (e.g. USGS)
    - Rock classification ontologies:
      - Multiple hierarchies (chemical, fabric, texture, genesis) from Geological Survey of Canada (GSC)
      - Single hierarchy from British Geological Survey (BGS)
- **Problem**:
  - Support **uniform queries** across all maps
  - … using **different** ontologies
  - Support **registration** w/ ontology A, **querying** w/ ontology B

Schema Integration
("registering" local schemas to the global schema)



Multihierarchical Rock Classification "Ontology" (Taxonomies) for "Thematic Queries" (GSC)

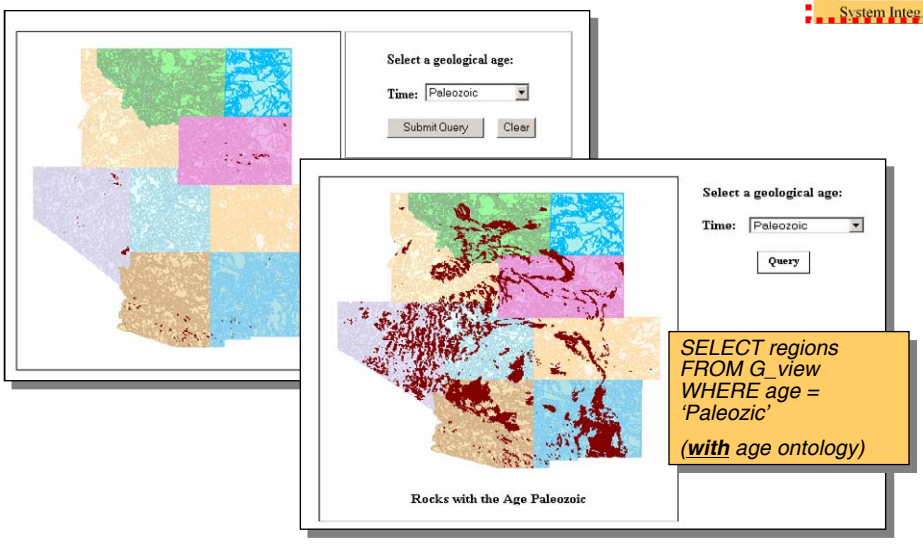
Synthesis
Semantics
Structure
Syntax
tem Integ.



SDM Tutorial, EDBT'06, Gertz, Ludäscher

Ontology-Enabled Application Example: **Geologic Map Integration**

Synthesis
Semantics
Structure
Syntax
System Integ.



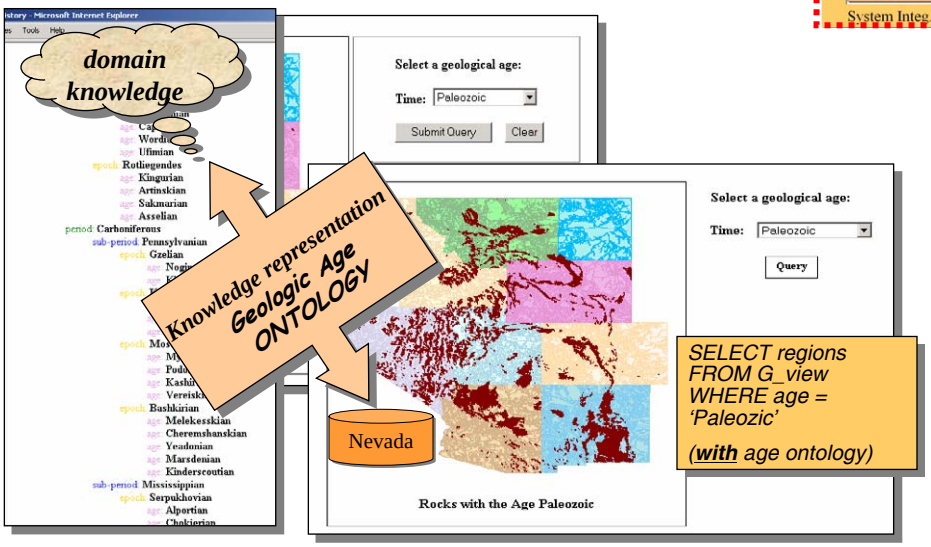
SELECT regions
FROM G_view
WHERE age = 'Paleozic'

*(**with** age ontology)*

SDM Tutorial, EDBT'06, Gertz, Ludäscher

Ontology-Enabled Application Example: **Geologic Map Integration**

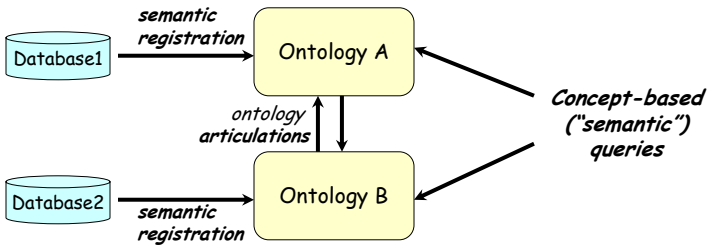
Synthesis
Semantics
Structure
Syntax
System Integ.



Semantic Mediation (via "semantic registration" of schemas and ontology articulations)

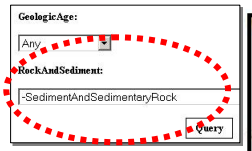
Synthesis
Semantics
Structure
Syntax
System Integ.

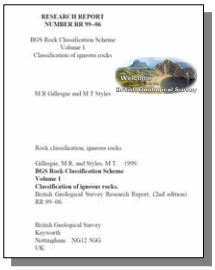
- **Schema elements** and/or data values are **associated** with concept expressions from the target **ontology**
  - ➔ conceptual queries "through" the ontology
- **Articulation ontology**
  - ➔ source registration to A, querying through B
- **Semantic mediation: query rewriting w/ ontologies**

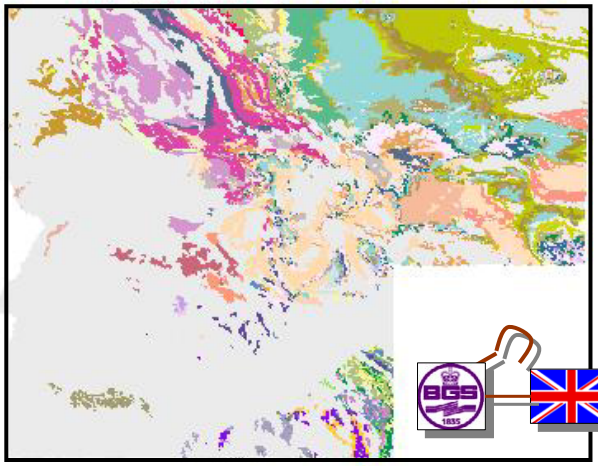


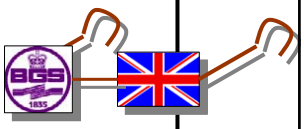
Sedimentary Rocks: BGS Ontology

Synthesis
Semantics
Structure
Syntax
System Integ.



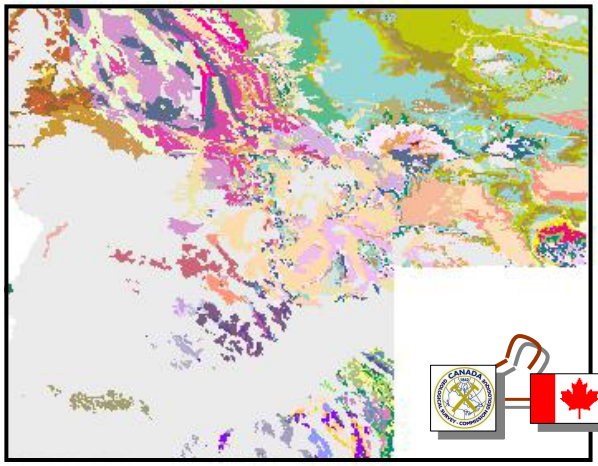


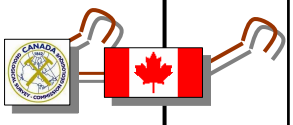




Sedimentary Rocks: GSC Ontology

Synthesis
Semantics
Structure
Syntax
System Integ.





Metadata

- **Metadata:**
  - Literally: "**data about data**"
  - Descriptive (meta-)information about the "actual" data
  - Borderline between data and metadata not always clear
  - "Someone's metadata is someone else's data …"
  - Related notions:
    - **schema, data dictionary, conceptual model, ontology, …**
  - Nevertheless, metadata is a useful concept to bundle a number of aspects of describing data
- **Traditional example: Library catalog card …**
  - contains "metadata" about the contents and location of books

Perspectives on Metadata

- **(Digital) Library Perspective: Metadata = Catalog Records**
  - ***Resources***: things that can be identified
    - documents, web pages, images, sound files, teaching packages, books, museum objects, people, organizations
  - ***Metadata***: structured information about resources
    - May be included with resources ("embedded/inlined metadata") or collected in separate "union catalogues"
    - Some from the resource itself (size, format), some from external sources (**provenance**, location, accessibility)

➔ **Usage aspects:**

- Classification … of (digital) holdings
- Search … of the (metadata) catalog for the …
- Retrieval … of (digital) library objects

Perspectives on Metadata

- **Database Perspective: Metadata = Schema Information**
  - In RDBMS: data dictionary (table structure, column names+types, …)
  - In XML: XML DTDs and XML Schema

➔ **Usage aspects:**

- Schemas describe the logical ("almost physical") structure of the data
- Knowledge of the schemas is necessary to articulate database queries
- Querying can be more powerful (and complex) than retrieval
  - querying as "computation" (e.g. selection … join,… transitive closure…shortest paths …)

Perspectives on Metadata

- **Conceptual Modeling Perspective:**
  - We can view an UML or ER diagram as metadata about data in a DBMS
  - While the logical database schema (table structure) has some resemblance with the original conceptual model, it might be "normalized" (➔ graduate course on database theory) for efficiency reasons or to avoid update anomalies
  - In contrast, a conceptual model (UML/ER) emphasizes the core entities (classes) and their relationships and properties, independent of a concrete database schema

➔ **Usage aspects:**

- To capture high-level conceptual ("semantic") information about the stored data; documentation
- Important during the design of a database / information system
- Current practice: this important form of "metadata" is often lost after the initial design phase ➔ attempts to re(verse-)engineer conceptual-level information
  - e.g. via ontologies and "semantic registration", i.e., mappings between logical schema and ontology-level information

Other Aspects of Metadata

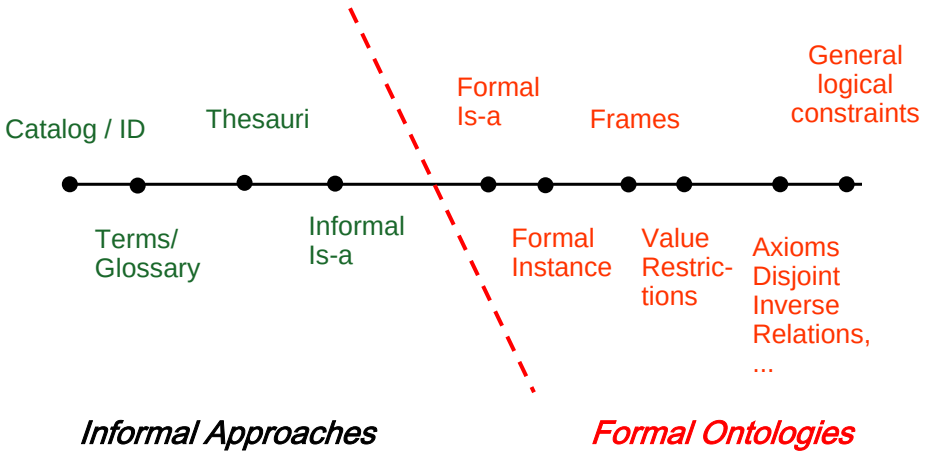
- **Interpretation of data**
  - Especially relevant for scientific applications
- **Reuse of data**
  - Human-readable or machine-processable metadata support to reuse data
- **Preservation of data**
  - Making data available and useful in the future (Rosetta Stone)
  - Special standards for preservation metadata
    - OAI (Open Archives Initiative): http://www.openarchives.org/

Metadata Standards

- **Content standards:**
  - which pieces of information are to be recorded (e.g. DC)
- **Value standards:**
  - how is the information to be recorded (= DC *encoding schemes)*
  - formats (ISO date format, NCA name formats, AACR)
  - lists of valid values (thesauri, authority files)
- **Structure standards:**
  - how the information is to be grouped and labelled for use by computers and humans (XML schemas, MARC)

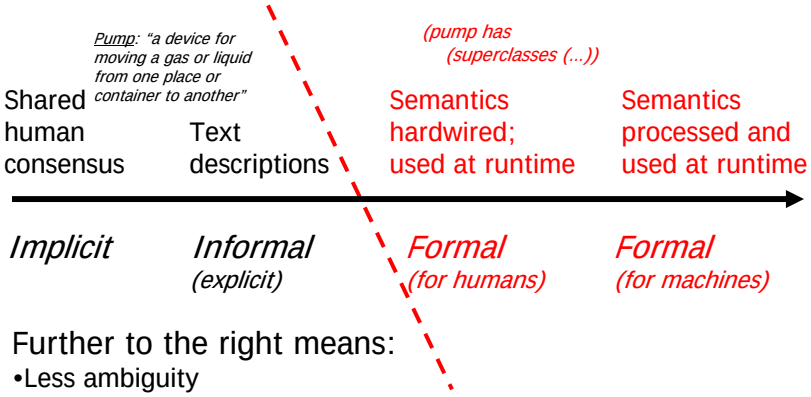
Src: Leonard Will

Ontologies and their Relatives



NOTE: Not always, more formal = better …

Ontologies: A Semantic Continuum



Further to the right means:

- Less ambiguity
- More likely to have correct functionality
- Better inter-operation (hopefully)
- Less hardwiring
- More robust to change
- **More difficult**

Src: [Mike Uschold, Boeing Corp]

Why do we need controlled vocabularies?

- **Need**: The ability to make **cross database/species queries** such as
  - ***"What mutations affect inflorescences in tomato, rice, maize and Arabidopsis?"***
- **Problem**: **Terminology** between databases and taxa is **heterogeneous**
  - e.g. an inflorescence can be called an ear, tassel, raceme, cyme, etc.
- **Solution**: Create ***controlled vocabularies*** that are used by databases to describe biological data from different organisms

Src: Plantontology.org

Gene Ontology http://www.geneontology.org

- **Gene Ontology (GO)**
  - most well-known example of a biological ontology
  - describes three biological domains:
    - sub-cellular localization,
    - molecular function (what the gene does) and
    - biological process (the cellular, developmental or physiological events the gene product is involved in)
  - concepts are used to describe gene products
    - descriptions are species independent

    ➔ meaningful and efficient comparisons of genes across diverse taxa are possible

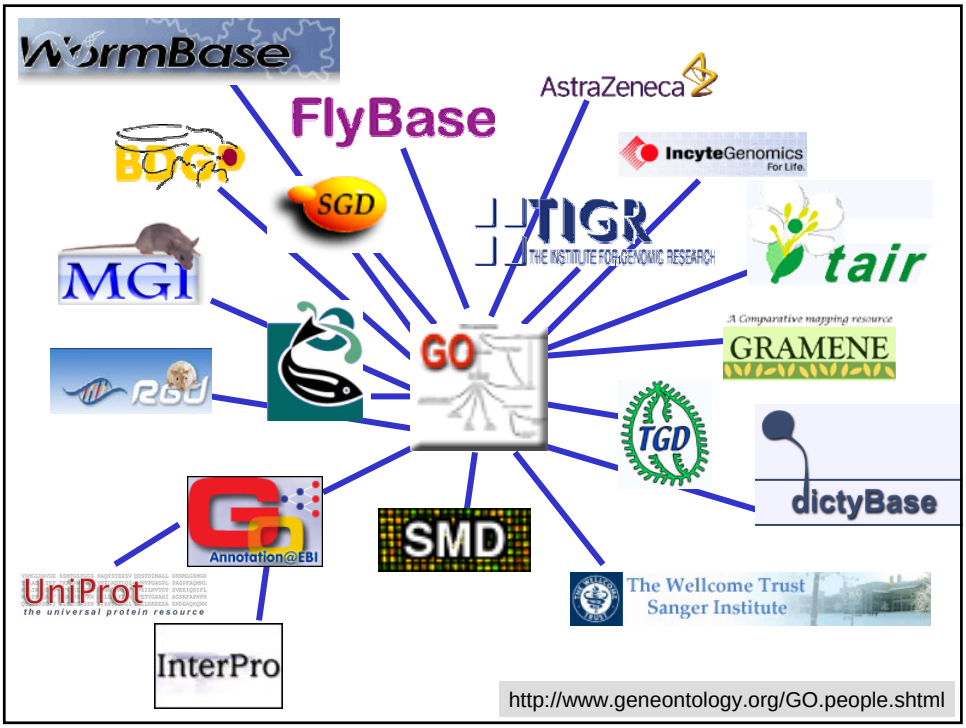
Src: Plantontology.org

How ontologies/controlled vocabularies help …

- Using GO terms to describe different aspects of a gene product it becomes possible to **find/group genes** by **biological concepts/functions** rather than being limited by sequence similarity or nomenclature

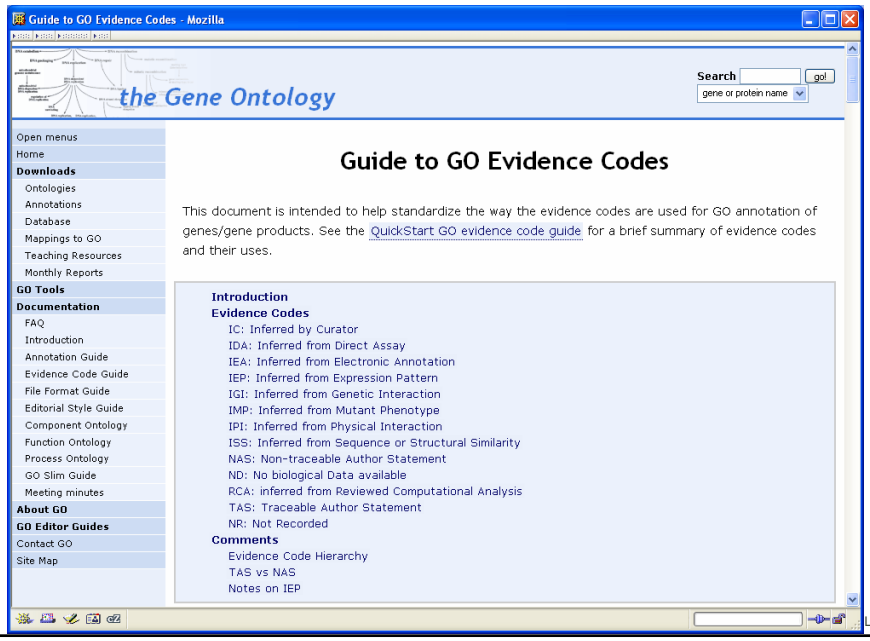
- Example:
  - Taking 4 kinases at random from 4 different organisms (Fly: ZWIM, Mouse: DAPK2 and Arabidopsis: KIPK and yeast: WEE1), based on gene nomenclature it is not clear that all of these genes encode protein kinases!
  - However, in GO, all 4 genes are annotated to the term "*protein kinase activity*" making it simple to find genes with similar functions in diverse species.

Src: Plantontology.org



http://www.geneontology.org/GO.people.shtml

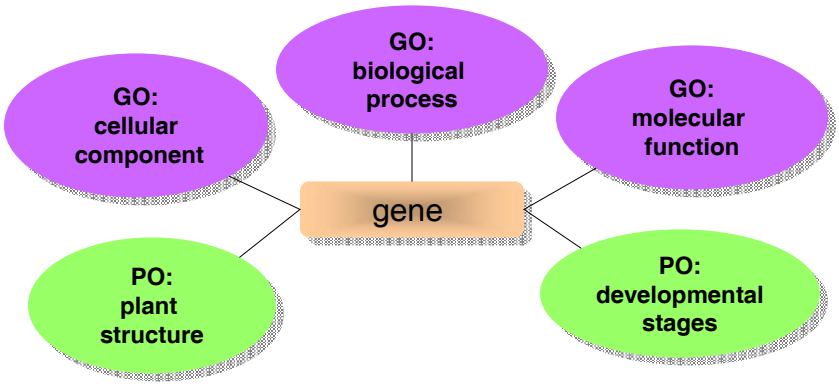
GO Evidence Codes



Ludäscher

The relationship of PO to other ontologies

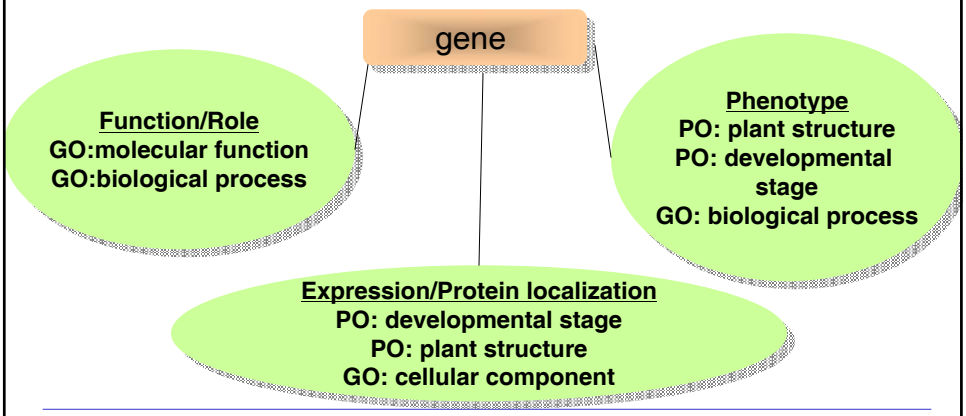
- **The plant ontologies complement and do not overlap with other ontologies that describe characteristics of gene products.**



Src: Plantontology.org

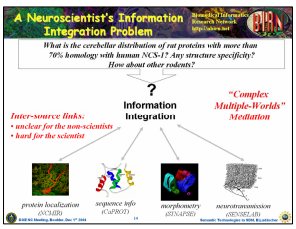
Using ontologies to describe gene products

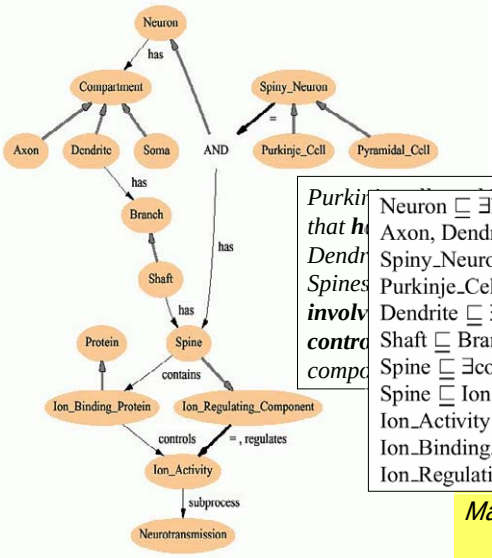
- **One or more ontologies can be used to describe different aspects of the gene product.**



Src: Plantontology.org

Formal Ontologies to "**glue**"
SYNAPSE & NCMIR Data





Neuron ⊑ ∃has.Compartment
Axon, Dendrite, Soma ⊑ Compartment
Spiny_Neuron ≡ Neuron ⊓ ∃has.Spine
Purkinje_Cell, Pyramidal_Cell ⊑ Spiny_Neuron
Dendrite ⊑ ∃has.Branch
Shaft ⊑ Branch ⊓ ∃has.Spine
Spine ⊑ ∃contains.Ion_Binding_Protein
Spine ⊑ Ion_Regulating_Component
Ion_Activity ⊑ ∃subprocess_of.Neurotransmission
Ion_Binding_Protein ⊑ Protein ⊓ ∃controls.Ion_Activity
Ion_Regulating_Component ≡ ∃regulates.Ion_Activity

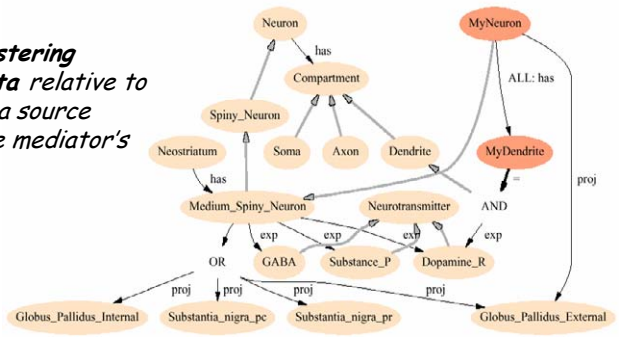
*Made usable for the **system** using Description Logic*

formalized as an "ontology graph"

Source **Data Contextualization** through **Ontology Refinement**

Biomedical Informatics Research Network

*In addition to **registering** ("hanging off") **data** relative to existing concepts, a source may also **refine** the mediator's domain map...*

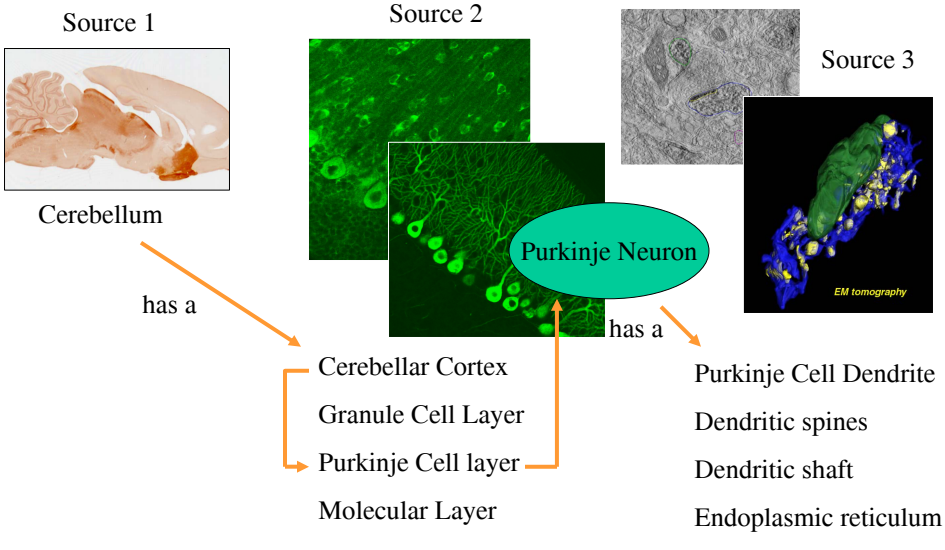


$$\text{MyDendrite} \equiv \text{Dendrite} \sqcap \exists \text{exp.Dopamine\_R}$$

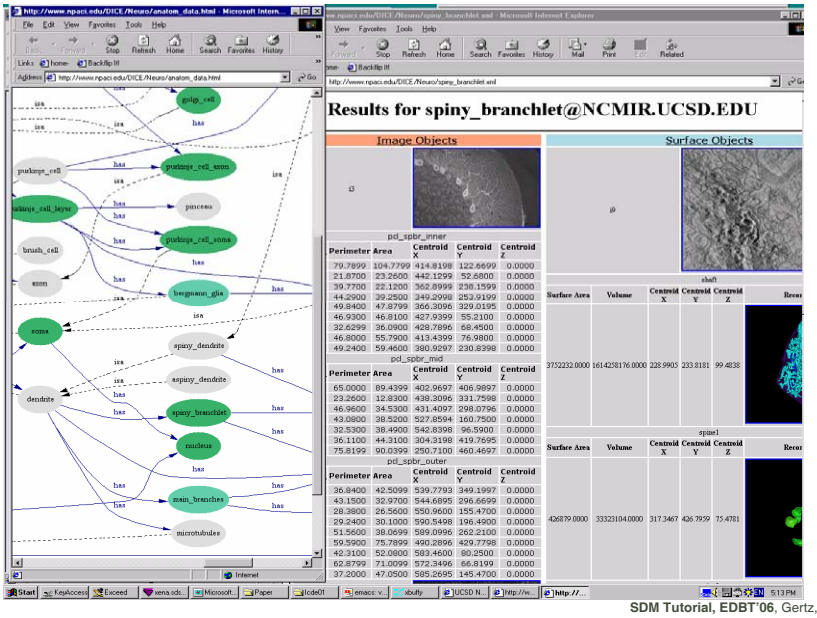
$$\text{MyNeuron} \sqsubseteq \text{Medium\_Spiny\_Neuron} \sqcap \exists \text{proj.Globus\_pallidus\_external} \sqcap \forall \text{has.MyDendrite}$$

⇒ *sources can register **new concepts** at the mediator ...*

Querying across Scales via "Glue" Knowledge

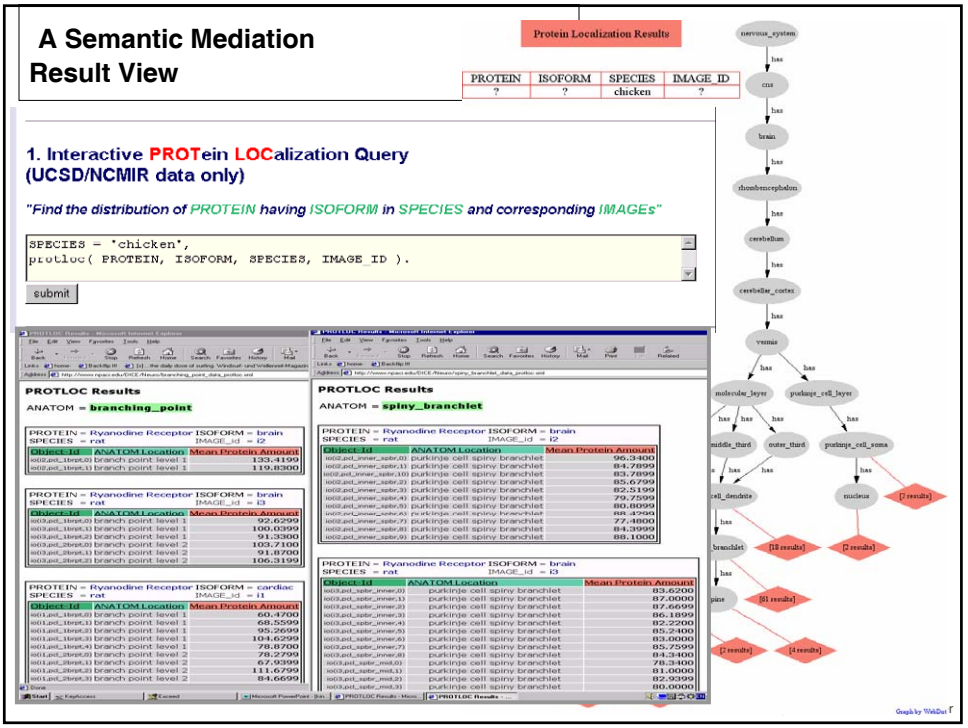


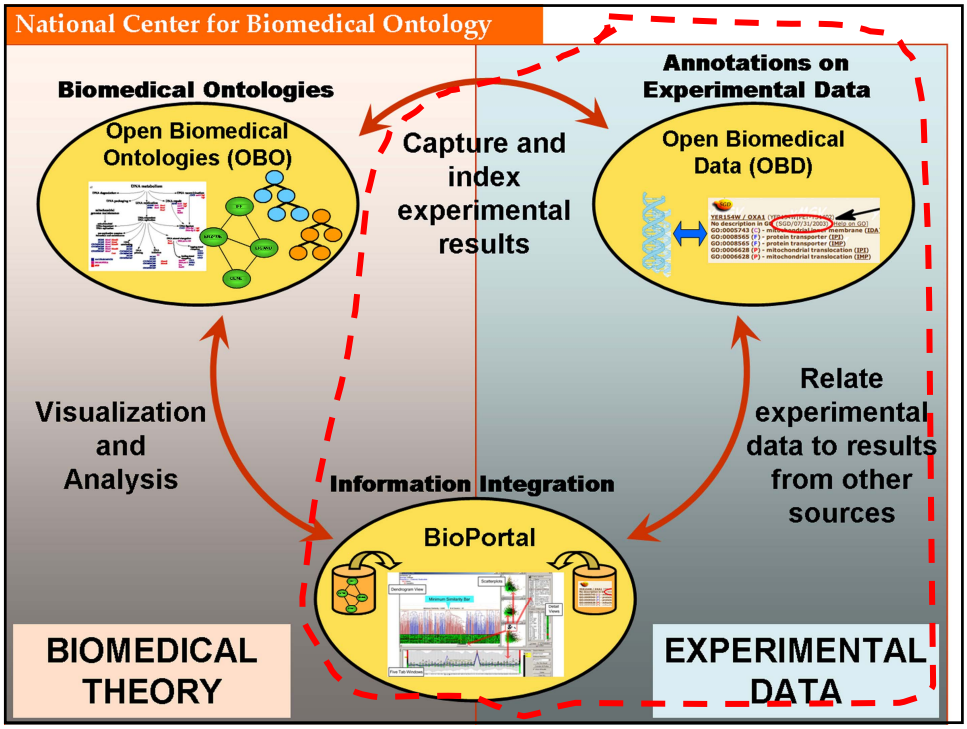
"Semantic Source Browsing":
Domain Maps/Ontologies (left) & conceptually linked data (right)

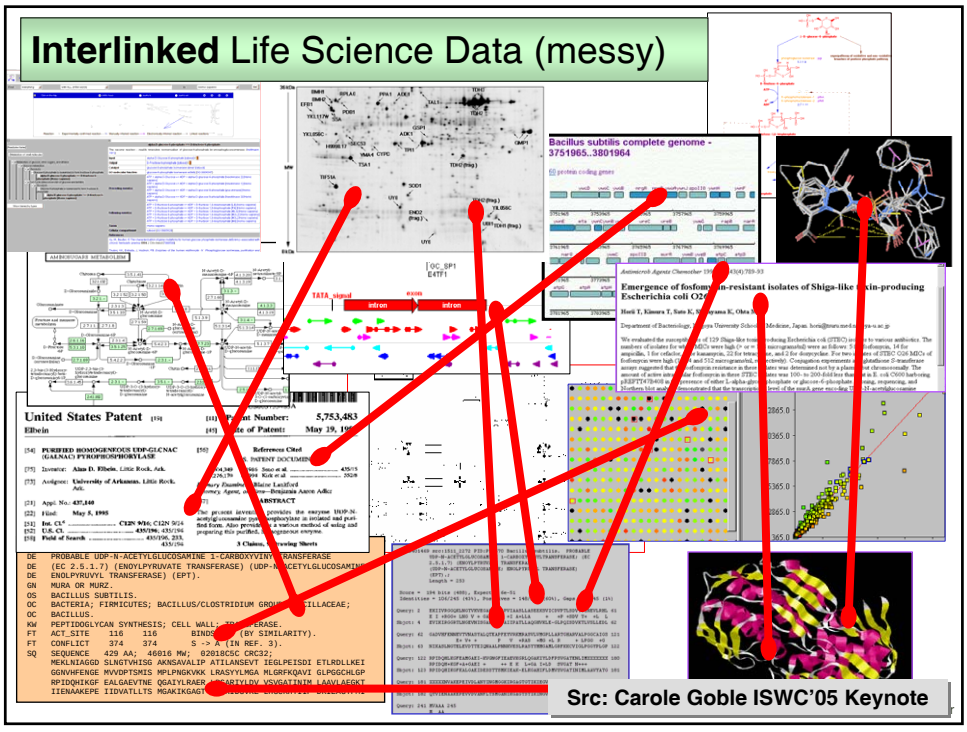


SDM Tutorial, EDBT'06, Gertz, Ludäscher

A Semantic Mediation Result View




National Center for Biomedical Ontology
Biomedical Ontologies
Open Biomedical Ontologies (OBO)
Capture and index experimental results
Annotations on Experimental Data
Open Biomedical Data (OBD)
Visualization and Analysis
Information Integration
BioPortal
Relate experimental data to results from other sources
BIOMEDICAL THEORY
EXPERIMENTAL DATA


Interlinked Life Science Data (messy)
Src: Carole Goble ISWC'05 Keynote

Two Cultures of Integration

- **Schema-driven (computer scientists/DB folks)**
  - Much smaller than data, (hopefully) well-defined elements
  - Resolve redundancy and heterogeneity at the schema level
  - High degree of automation once system is set-up
  - Focus on methods - you rarely publish a "data paper"
- **Data-driven (biologists)**
  - Value is in the data, abstraction is a result of analysis
  - Don't bother with schemas
    - Abstraction is volatile and depends on experimental technique
  - Manual integration at data level, constant high effort
  - You rarely publish a (database) "method paper"

Src: Ulf Leser, Felix Bauman, CIDR 2005

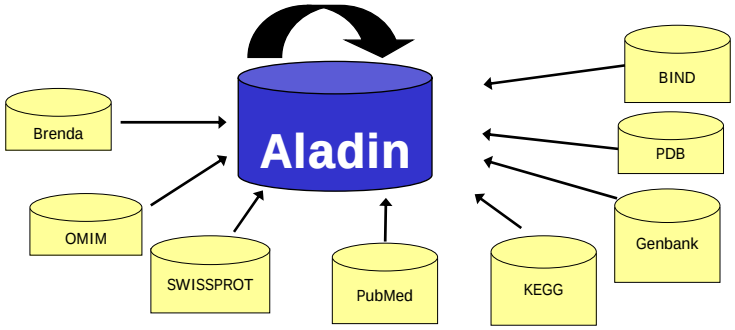
- **Component+Process-driven (engineers)**
  - Model, design, and build efficient, large and/or complex systems
  - Focus on models of computation, interaction, simulation
  - Publish method and systems papers

  ➔ Module IV

SDM Tutorial, EDBT'06, Gertz, Ludäscher

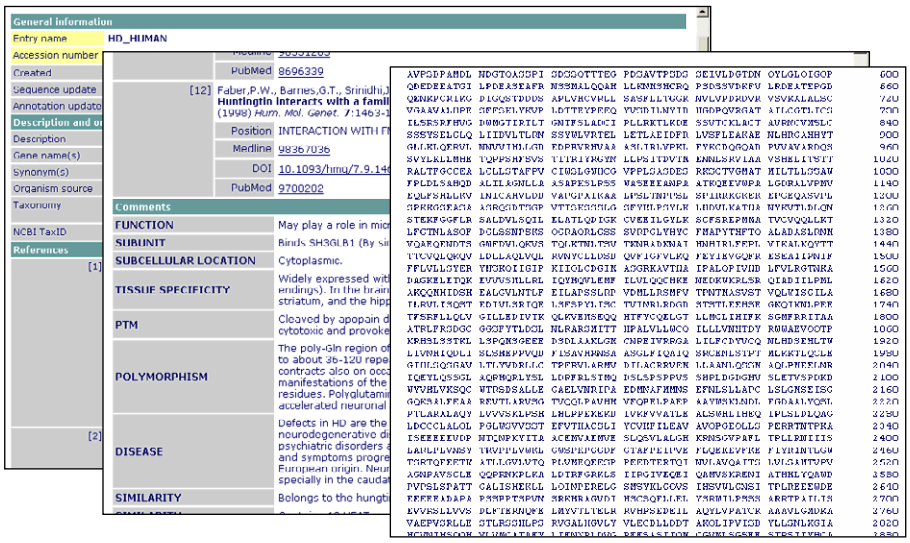
Example: Aladin (Leser & Naumann)

- **ALmost Automatic Data INtegration for the Life Sciences**
  - Minimize manual effort
  - Keep quality of integrated data as high as possible
  - Use domain-specific heuristics



Src: Ulf Leser, Felix Nauman, CIDR 2005

A Biological Database (e.g. UniProt)

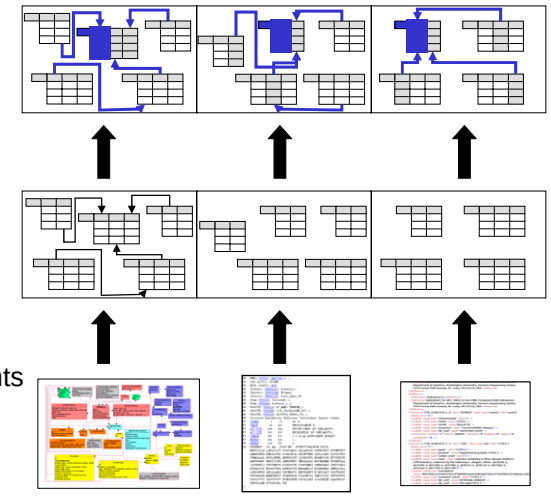


Src: Ulf Leser, Felix Bauman, CIDR 2005

Aladin: Steps 1, 2, 3

Steps 2 and 3
- Guess primary objects
- Guess accession number
- Guess / find FK constraints

Step 1
- Parse and import
- Arbitrary target schema
- With or without FK constraints

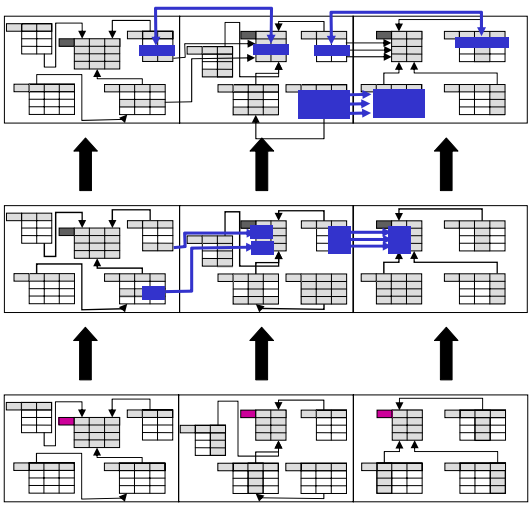


Src: Ulf Leser, Felix Bauman, CIDR 2005

Aladin: Steps 4, 5

Step 5
- Guess duplicates
- Different degrees of "duplicateness"

Step 4
- Guess existing cross-refs
- Compute new cross-refs



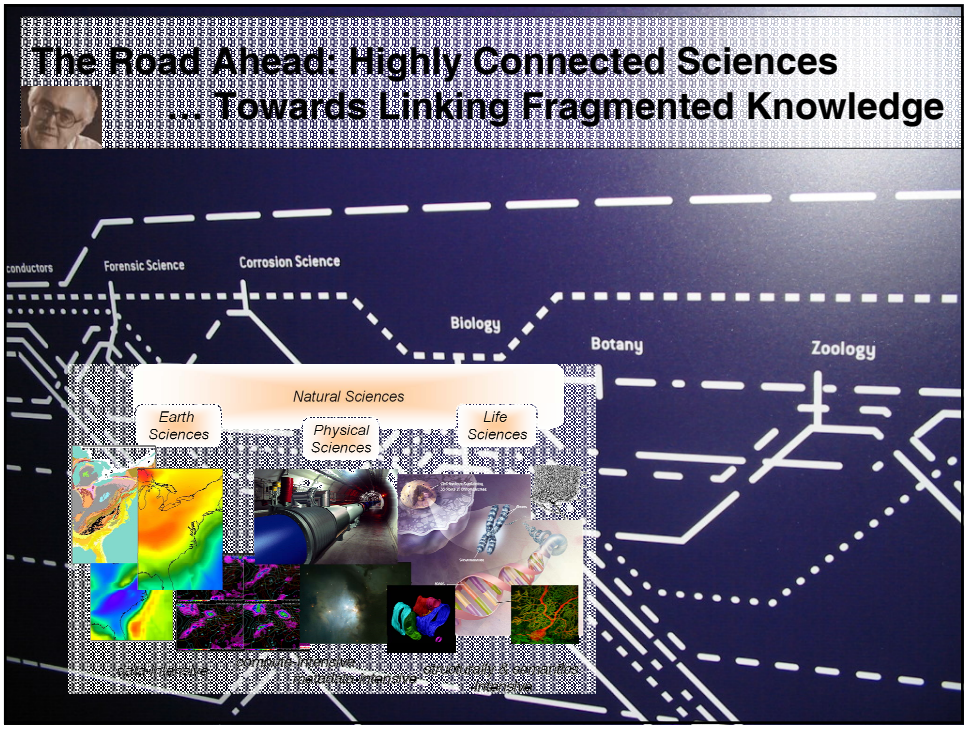
Src: Ulf Leser, Felix Bauman, CIDR 2005

Caveats

- **Not meant for high-throughput data**
  - Proteomics profiling, gene expression databases
  - Targets "knowledge-rich" databases
- **Resulting warehouse will contain errors**
  - Wrong cross-refs, misinterpreted structure, missing links
  - Requirement: Measure quality of Aladin's methods
    - Use existing integrated databases as gold standard
    - Precision/recall measures can be derived for all steps
- **Intended for human usage, not for automatic further processing**

Src: Ulf Leser, Felix Bauman, CIDR 2005

The Road Ahead: Highly Connected Sciences … Towards Linking Fragmented Knowledge



Summary: From DI to Scientific DI

- **Scientific data management is at the heart of CI and e-Science**
- **Scientific DI = conventional DI + KR + Link-based Integration + …**
- **"A little semantics goes a long way" (SemWeb folks)**
- **"A little database technology goes a long way" (DB folks)**

… need both … and more!