Semantic Web Technologies

Lecture 8: SWT for the Life Sciences 1: Background and data integration

Maria Keet
email: keet -AT- inf.unibz.it
home: http://www.meteck.org
blog:
http://keet.wordpress.com/category/computer-science/72010-semwebtech/

KRDB Research Centre
Free University of Bozen-Bolzano, Italy

15 December 2009

Outline

Ontologies to solve real problems
- 'Historical' overview from GO to OBO Foundry
- Late early adopters

Linking Data
- Data integration strategies
- Linked data using ontologies

Linking technologies
- Preliminary points
- SWT for SWLS

Early bioinformatics

- Advances in technologies to sequence genomes in the late '80s-early'90s, as well as more technologies for proteins
- Need to store the data: in databases ('90s)
- Several 'model organism' databases with genes (and genomes) of the fruitfly, yeast, mouse, a flowering plant, flatworm, zebrafish
- Compare genes and genomes
  - One observation (of many): About 12% (some 18,000) of the worm genes encode proteins whose biological roles could be inferred from their similarity to their (putative) orthologues in yeast, comprising about 27% of the yeast genes (about 5,700)
  - *What else can we infer from comparing genes and genomes (across species)?*
  - *What about the possibility of automated transfer of biological annotations from the model organisms to less 'fancy' organisms based on gene and protein sequence similarity, to use to improve human health or agriculture?*

Scope and requirements

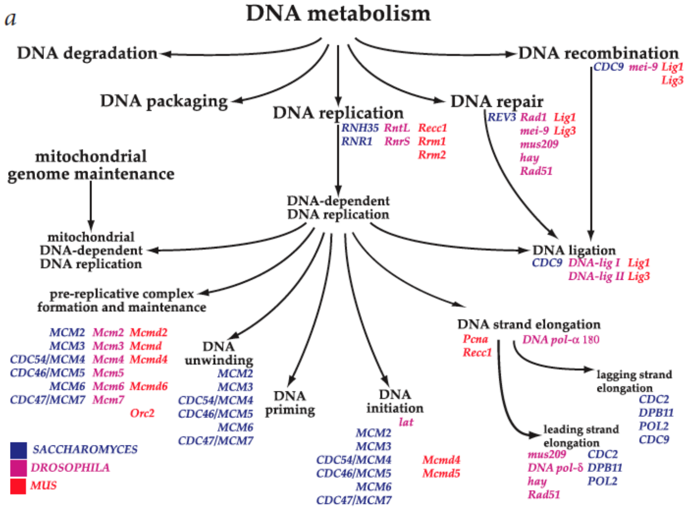
- Need: a mainly computational system for comparing or transferring annotation among different species
- Methods for sequence comparison existed
- Main requirements:
  - One needs a shared, controlled, vocabulary for annotation of the gene products, the location where they are active, the function they perform
  - To take on board and be compatible with existing terminologies, like gene and protein keyword databases such as UniProt, GenBank, Pfam, ENZYME etc.
  - Database interoperability among, at least, the model organism databases
  - Organize, describe, query and visualize biological knowledge at vastly different stages of completeness
  - Any system must be flexible and tolerant of this constantly changing level of knowledge and allow updates on a continuing basis

How to meet such requirements?

- Two main strands in activities:
  - Very early adopters from late 1990s (by sub-cellular bio), i.e., starting *without* Semantic Web Technologies
  - Early adopters from mid 2000s (e.g., eco), starting *with* Semantic Web Technologies
- The Gene Ontology Consortium
  - Initiated by fly, yeast and mouse database curators[1] and others came on board (see `http://ww.geneontology.org` for a full list)
  - In the beginning, there was the flat file format .obo to store the ontologies, definitions of terms and gene associations
  - Several techniques on offer for data(base) integration that could be experimented with

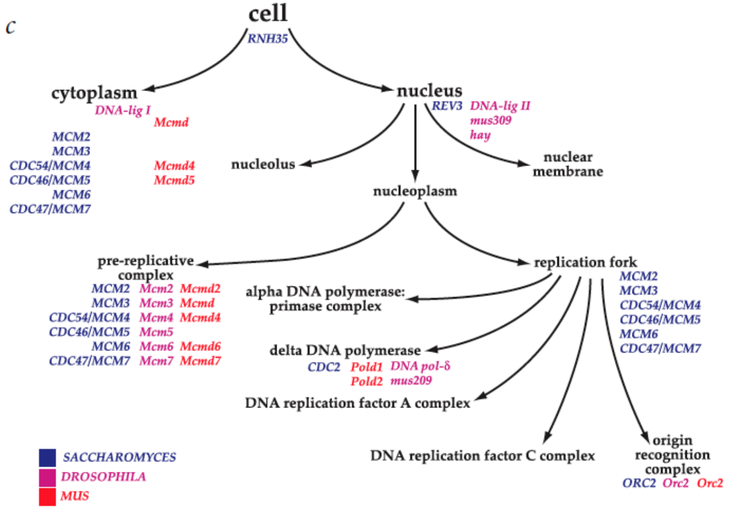
[1] more precisely: FlyBase (http://www.flybase.bio.indiana.edu), Berkeley Drosophila Genome Project (http://fruitfly.bdgp.berkeley.edu), Saccharomyces Genome Database (http://genome-www.stanford.edu), and Mouse Genome Database and Gene Expression Database (http://www.informatics.jax.org).

GO contents example (process)



from GOC, 2000

GO contents example (cellular component)



from GOC, 2000

Progress

- Tool development, e.g. to:
  - add and query its contents
  - annotate genes (semi-automatically)
  - link the three GO ontologies
  - mine the literature (NLP)
- Content development: more in the GO, extensions to the GO (e.g., rice traits), copy of the principle to other subject domains (e.g., zebrafish anatomy)
- The GO and its approach went well beyond the initial scope (which does not imply that all requirements were met fully)

Toward an update of the approach and contents

- Problems:
  - one can infer very little knowledge from the obo-based bio-ontologies (mainly where there are errors, but not *new* insights)—but note that that was not it's original aim
  - semantics of the relations overloaded
  - mushrooming of obo-based bio-ontologies by different communities, which makes interoperation of the ontologies difficult
  - greater needs for collaborative ontology development, maintenance, etc.
- Proposed solution: structured, coordinated, development of ontologies adhering to a set of principles: the OBO Foundry

OBO Foundry

- Extending the **O**pen **B**iological **O**ntologies principles...
  - open,
  - orthogonal,
  - same syntax,
  - common space for identifiers
- ... to one for the **O**pen **B**iomedical **O**ntologies:
  - developed in a collaborative effort
  - usage of common relations that are unambiguously defined (*in casu*: the Relation Ontology)
  - provide procedures for user feedback and for identifying successive versions
  - has to have a clearly bounded subject-matter ("so that an ontology devoted to cell components, for example, should not include terms like 'database' or 'integer' " ...)

More info in Smith et al, 2007, and http://www.obofoundry.org

OBO Foundry

- Sorting out the ontologies we have; e.g.,
  - harmonizing the four cell type ontologies into one (CL)
  - coordinating the anatomy ontologies of the various model organisms through a Common Aanatomy Reference Ontology
  - modularization of the subject domain by granularity, continuants, and occurents
- Adding ontologies to fill the gaps
- making OBO and OWL ontologies interoperable
- "Our long-term goal is that the data generated through biomedical research should form a single, consistent, cumulatively expanding and algorithmically tractable whole"
- "The result is an expanding family of ontologies designed to be interoperable and logically well formed and to incorporate accurate representations of biological reality"
- Aimed at "coordinated evolution of ontologies to support biomedical data integration"

Other early adopters of bio-ontologies

- Start with a 'clean slate': ontology engineering straight into OWL, e.g.:
  - Ontologies in ecology (Madin et al, 2008)
  - Biopax, who are now going into two directions: one as ontology-as-scientific-theory and another version as ontology-for-applications (see http://www.biopax.org))
  - protein phosphatases (Wolstencroft et al, 2007)
- Linking external data to the ontologies, e.g.:
  - HistOn ontology (in OWL) and an RDF triple store with Sesame (Marshall et al, 2006)
  - Ontology-Based Data Access case study with HGT 'application ontology' in roughly OWL 2 QL and data in an RDBMS (Calvanese et al, 2010)

Aims

- Not just for data integration
- More precise and accurate representation of knowledge/reality (than with obo format, SKOS etc.)
- Aim also to do *automated reasoning* over it; e.g.:
  - instance classification
  - hypothesis testing
  - intelligent access to the data by using terms in the ontology instead of the gory details of the database
  - more sophisticated ontology browsing

Shopping for approaches to achieve data integration

I. Physical schema mappings
  - Global As View (GAV)
  - Local As View (LAV)
  - GLAV

II. Conceptual model-based data integration

III. Data federation

IV. Data warehouses

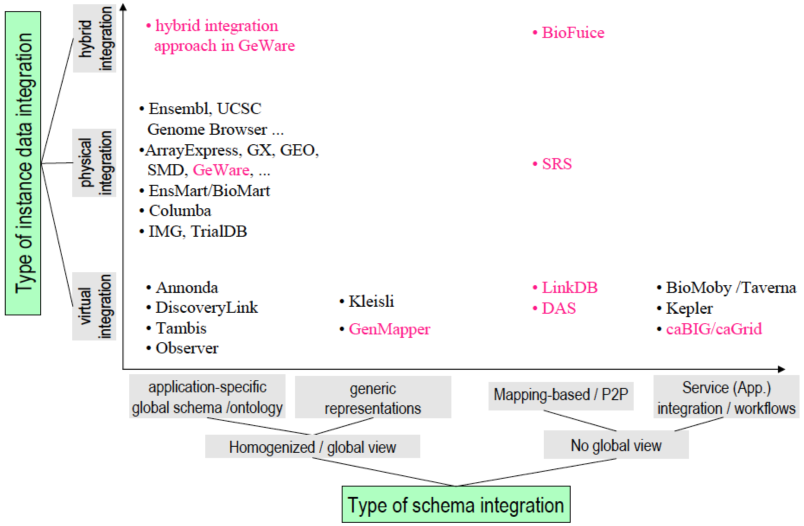
V. Data marts

VI. Services-mediated integration

VII. Peer-to-peer data integration

VIII. Ontology-based data integration
  - I or II (possibly in conjunction with the others) through an ontology
  - Linked data by means of an ontology

Classification of data integration approaches and tools



Type of instance data integration

Type of schema integration: Homogenized / global view; No global view

	application-specific global schema /ontology	generic representations	Mapping-based / P2P	Service (App.) integration / workflows
hybrid integration	• hybrid integration approach in GeWare		• BioFuice	
physical integration	• Ensembl, UCSC Genome Browser ... •ArrayExpress, GX, GEO, SMD, GeWare, ... • EnsMart/BioMart • Columba • IMG, TrialDB		• SRS	
virtual integration	• Annonda • DiscoveryLink • Tambis • Observer	• Kleisli • GenMapper	• LinkDB • DAS	• BioMoby /Taverna • Kepler • caBIG/caGrid

E. Rahm et al.: Data Integration in Bioinformatics and Life Sciences EDBT summer school 2007 21

Overview

- Ontology on top of physical schemas?
- Ontology on top of conceptual data models
- Ontology to mediate between services
- Classifying instances into an ontology

Linked data in Bio

- *Data-level integration*
- Annotated instances stored in databases
- Across databases at physically different locations
- On the Web
- Where the ontology tells you which ones are the same, or instantiating the same universal represented in the ontology

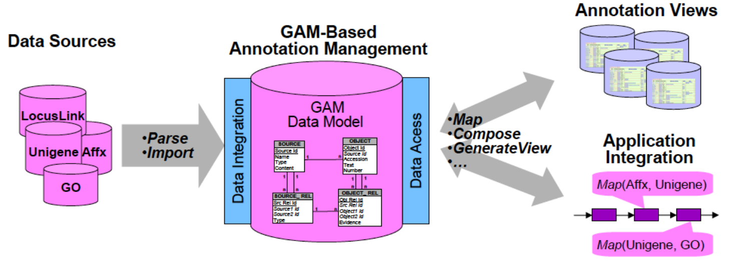
Web-links based ‘integration’

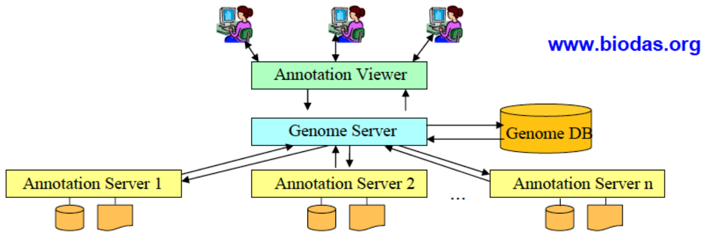
- Web-Link = URL of a source + ID of the object of interest
- Little integration effort, Scaleable, Navigational analysis: only *one object at a time*
- A mere link is semantics-poor w.r.t. language and subject domain meaning, e.g.:
  - How would one do automated reasoning with it to derive implicit knowledge? (not)
  - “related to” versus, among others, *partOf*, *isA*, *containedIn* etc; i.e., even poorer than the thesaurus’ RT, BT, NT
- DBGET + LinkDB
- see also `http://www.genome.jp/dbget/`

Integration and annotations examples

- GenMapper
  - Centralised, with a global view
  - Exploits existing mappings between objects/sources
  - Links between the databases through the annotations of the objects (e.g., genes, proteins)
  - Links to terms of the ontology (GO), i.e., (semi-)*manual* classification
- Distributed Annotation Systems (DAS)
  - Distributed, mapping-based, no global view
  - Central genome server as primary source that contains the reference genome sequence
  - Separately, several annotation servers where the sources are wrapped
  - Recalculation of all annotations when the reference sequence has changed

Integration and annotations examples





from Rahm et al, 2007

Other integration systems (examples)

- BioFuice, based on iFuice:
  - Use instance-level cross-references for instance-level mappings between sources
  - Mappings have a semantic mapping type
  - Domain model (± an ontology) indicates available object types and relationships
- Sequence Retrieval System: wrapping sources, making them accessible through one interface
- BioGuide: selecting appropriate sources and tools using chosen preferences and strategy
- IMGT-Choreography based on the IMGT-ONTOLOGY concepts to coordinate services among databases
- Mash-ups, RDF, XML, ...

Generalising the current bio-integration implementations

- Many CS theory and technologies 'on offer' that purport to solve each integration problem
- All of them experimented with by the users, who added linked data, annotations, and web-links to the array of options
- For all: still a lot of manual work
- For all: for various reasons fairly simple end-user level queries (which might well be complicated at the back-end)
- Does it actually solve the original problem and address the requirements as defined by the GOC? (see slide 6)
- Ontology usage: 'simple' ontologies, or none at all
- Semantic Web Technologies usage: ...

Background

- Main players in SWLS are engineers, domain experts, bioinformaticians, bio-ontologists. "Something bio" covers many disciplines: e.g., genomics, metabolomics, ecoinformatics, and, above all: biomed & healthcare. Diverse fields, diverse needs.
- Some current characteristics:
  - Collaboration & interdisciplinary work
  - Possible not-intended use of technologies (from the perspective of computer scientist)
  - Novel-ness of the technologies: data integration techniques of the '90s did not solve the issues, SW tech will?
  - Goal-driven: looking for the "killer app" and discover novel information about nature.
  - Thus far, there are very few success stories

Expressive ontology vs scalability & performance

- Some ontologies in OWL (2007), denoted in their DL language used

Ontology	Characterizing DL
ProPreO	$\mathcal{SHOIN}(D)$
BioPAX	$\mathcal{ALCHON}(D)$
Cell Cycle Ontology	$\mathcal{SIN}(\mathcal{D})$
HistOn	$\mathcal{ALCHIF}(D)$
NMR Ontology	$\mathcal{SHF}$
MGED Ontology	$\mathcal{ALEOF}(D)$
Human Developmental Anatomy Ontology	$\mathcal{ALEOF}(D)$
Microbial Loop	$\mathcal{ALCHI}$
Gene Ontology	$\mathcal{ALE}(D)$
Protein-Protein Interaction Ontology	$\mathcal{ALE}(D)$
Mammalian Phenotype Ontology	$\mathcal{AL}(D)$
Disease Ontology	$\mathcal{AL}$
FungalWeb	$\mathcal{FL}_0$

- "Breakpoint" is known roughly and through disparate experiments, but not (yet) through benchmarking
- Lite-izing ontologies

Queries in the SW

- What can you do? We have:
  - Within the SW-scope, we have: SPARQL, SeRQL, Sesame, XQuery, XPath, Xcerpt, Prova, ...
  - Know their strengths and weaknesses[2], tool support
  - Performance issues (e.g. interval join with several query languages Cell Cycle Ontolog browsing)
- But is that what the user wants?
  - Recursive queries
  - Subgraph isomorphisms
  - Query data through the ontology
  - Traverse paths of arbitrary (finite, but not pre-defined) length
  - ...

[2] e.g., Royer, L., Linse, B., Wächter, T., Furch, T., Bry, F., Schroeder, M. Querying Semantic Web contents. In: *Semantic Web: revolutionizing knowledge discovery in the life sciences*, Baker, C.J.O., Cheung, H. (eds), Springer: New York, 2007, 31-52

Examples

- D2RQ `http://sites.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/`: access the content of non-RDF databases, query with RDQL, SPARQL.
- A D2RQ graph wraps one or more local relational databases into a virtual, read-only RDF graph (Mappings between relational database schemata and OWL/RDFS ontologies). It rewrites Jena API calls, find() and RDQL queries to SQL queries and query answer is transformed into RDF triples that are passed up to Jena.
  - Non-bio example at `http://sites.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/#example`, and a bio-example in the BMC 2007 article
- From scratch: TFBS data → RDF → Sesame repository and query with SeRQL-S. Interval join with SeRQL (including SPARQL equivalent). see `http://integrativebioinformatics.nl/semanticdataintegration.html`

Semantic Web Technologies for HC & LS

- The Semantic Web will solve all your problems?
- W3C Health Care and Life Sciences Interest Group
  - "is designed to improve collaboration, research and development, and innovation adoption in the health care and life science industries. Aiding decision-making in clinical research, Semantic Web technologies will **bridge many forms of biological and medical information across institutions**."
  - "Subgroups focus on making biomedical data available in RDF, working with biomedical ontologies, prototyping clinical decision support systems, working on drug safety and efficacy communication, and supporting disease researchers navigating and annotating the large amount of potentially relevant literature."
- Example activity resulting in the BMC Bioinformatics articles "Advancing translational research with the Semantic Web" (2007) and "A journey to Semantic Web query federation in the life sciences" (2009)

From bench to bedside — and from CS theory to software application

Overview 23-author article by Ruttenberg et al, 2007

- "A significant barrier to translational research is the lack of uniformly structured data across related biomedical domains."
- "Current tools and standards are already adequate to implement components of the bench-to-bedside vision."
- "Gaps in standards and implementations still exist and adoption is limited by typical problems with early technology... growing pains as the technology is scaled up."
- SW "will improve the productivity of research, help raise the quality of health care, and enable scientists to formulate new hypotheses inspiring research based on clinical experiences"

What do they want?

- Data integration
- Querying the data across databases
- Expressive ontology languages to represent biological knowledge
- Manage (query) the data silos ('write-only database')
- Building upon the web of data
- Automation to 'upgrade' 'legacy' material to SemWeb technologies and standards
- Navigate and annotate potentially relevant literature

How do they do it?

- Global scope of identifiers
- RDFS/OWL
- Bottom-up development
  - RDF triple stores from 'legacy' RDBMS
  - Previously discussed bottom-up techniques for ontology development
- SWRL for rules

A few discussion questions

- Are (should?) "Tools and strategies to extract or translate from non-RDF data sources to enable their interoperability with data organized as statements." (be) part of the set of SW Technologies?
  - Or: where are (W3C) standardization efforts for RDBMS→RDF, excel→RDF, OBO→OWL, structured flat file → language y mappings?
- "BioRDF has the goal of converting a number of publicly available life sciences data sources into RDF and OWL."
  - Thus: *not using* SW Tech but *preparing for use*

A few discussion questions

- "While the need to integrate more types of data will continue, RDFS and OWL offer some relief to the burden of understanding data schemas."
  - Since when are ontologies read in their OWL syntax-format (or XML-serialised) human understandable? Did you learn RDFS on a rainy Sunday afternoon?
  - UML, ER, ORM, and conceptual graphs are well-established *graphical and formal* conceptual data modelling languages, is something wrong with using those ones?

A few discussion questions

- "A goal of the HCLSIG is to facilitate creation, evaluation and maintenance of core vocabularies and ontologies to support cross-community data integration and collaborative efforts. Although there has been substantial effort in recent years to tackle these problems, the methodology, tools, and strategies are not widely known to biomedical researchers."
  - Which "methodology, tools, and strategies"?
  - How would you address the lack of necessary skills of the (presumably intended) user-base of biomedical researchers?
- "The role of the ontologies task force is to work on well-defined use cases, supporting the other HCLSIG working groups."

A few discussion questions

Adaptable clinical pathways and protocols (ACPP)

- "The ACPP task force explores the use of Semantic Web technologies, including RDF, OWL, logic programming, and rules to represent clinical guidelines and guide their local adaptation and execution. ...*Representation of temporal concepts and inference rules necessary for tracking processes and ensuring temporal constraints on treatment.*"
  - How can one represent temporal concepts and constraint in RDF, OWL, Logic Programming or rules?
  - E.g. in OWL through a cumbersome reification and relate it to datatypes, time ontology in OWL, DL-Lite with role values, $\mathcal{DLR}_{US}$

A few discussion questions

- D2RQ "The mappings allow RDF applications to access the contents of relational databases using Semantic Web query languages like SPARQL. Doing such a mapping requires us to choose how tables, columns, and values in the database map to URIs for classes, properties, instances, and data values."
  - Name the pros and cons of RDF applications vs RDBMSs

Current identified technical limitations

- As listed in the article:
  - Scarcity of semantically annotated information sources
  - Performance and scalability
  - Representation of evidence and data provenance
  - Lack of a standard rule language
- Did you spot other limitations?

Summary

Ontologies to solve real problems

- ‘Historical’ overview from GO to OBO Foundry
- Late early adopters

Linking Data

- Data integration strategies
- Linked data using ontologies

Linking technologies

- Preliminary points
- SWT for SWLS