Semantic Web for the Life Sciences

Marijke Keet

KRDB Research Centre, Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
keet@inf.unibz.it

Semantic Web Technologies lecture d.d. May 23, 2007

Preliminary notes: no "new" SW technologies will be introduced in this lecture, but you will have to relate and compare technologies you have learned in previous lectures and *other courses.*

There is no lab associated with this lecture. Exercises are integrated in the "Case studies" section.

Outline

Background

- disclaimer....
- 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. [K05a]
- 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

General aspects of ontology development

- Developing bio-ontologies: mine legacy models in biology (STELLA [K05b], PathwayAssist), mine scientific literature, .obo format (DAGs) conversion to OWL, and *de novo*
- Structured approach with set of principles (OBO Foundry [9], RO [Setal05], foundational ontologies such as BFO [1], DOLCE [5], GFO [8]) and relatively centralised...
- ... versus customary freedom and development by smaller communities
- KR in which language? Develop with which tool (Protégé, SWOOP, GrOWL, Eclipse, ...)? Store in which format (XML serialised, RDBMS)?

Ontology languages

- Does it make sense to develop ontologies in RDF(S)?
- OWL-Lite, OWL-DL, WSMO etc., DL-Lite family [Cetal07]. OWL 1.1 "full" (based on $\mathcal{SROIQ}$ [HKS06]) and several candidates for OWL 1.1 "tractable fragment" [10].
- The subject domain is complex: choose most expressive language, **but**:
  - Bio-ontologies can become very large
  - Linking ontologies to data (wet-lab, medical records)
  - At least some users are 'spoilt' and demand good performance
  - Do they really "need" more expressive ontology languages? (Do you as logician decide that for them?)

A comparison

Language ⇒	OWL			*DL-Lite*			$\mathcal{DLR}$		
Feature ⇓	**Lite**	**DL**	**v1.1**	$\mathcal{F}$	$\mathcal{R}$	$\mathcal{A}$	*ifd*	μ	*reg*
Role hierarchy	+	+	+	-	+	+	+	+	+
N-ary roles (where $n \geq 2$)	-	-	-	±	±	±	+	+	+
Role concatenation	-	-	+	-	-	-	-	-	+
Role acyclicity	-	-	-	-	-	-	-	+	-
Symmetry	+	+	+	-	+	+	-	-	-
Role values	-	-	-	-	-	+	-	-	-
Qualified number restrictions	-	-	+	-	-	-	+	+	+
One-of, enumerated classes	-	+	+	-	-	-	-	-	-
Functional dependency	+	+	+	+	-	+	+	-	+
Covering constraint over concepts	-	+	+	-	-	-	+	+	+
Complement of concepts	-	+	+	+	+	+	+	+	+
Complement of roles	-	-	+	+	+	+	+	+	+
Concept identification	-	-	-	-	-	-	+	-	-
Range typing	-	+	+	-	+	+	+	+	+
Reflexivity *	-	-	+	-	-	-	-	+	+
Antisymmetry *	-	-	-	-	-	-	-	-	-
Transitivity * ‡	+	+	+	-	-	-	-	+	+
Asymmetry ‡	+	+	+	-	+	+	-	±	-
Irreflexivity ‡	-	-	+	-	-	-	-	+	-

Table: Differences between DL-based ontology and conceptual modelling languages (after [KR07]).

Expressive ontology vs scalability & performance

- OWL-DL is NExpTime-complete, $\mathcal{SROIQ}$ decidable, DL-LiteA PTime (logspace w.r.t. data complexity). [10]

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	$\mathcal{FL}_0$

Table: (after [KR07])

- "Breakpoint" is known roughly and through disparate experiments, but not (yet) through benchmarking
- Ontology integration or linking, and modularization
- 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 (e.g. [RLetal07]), tool support
  - Performance issues (e.g. interval join with several query languages [MPRB06] CCO browsing [2])
- 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
  - ...

Examples

- D2RQ [3]: 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 article [Retal07]
- From scratch [12]: TFBS data → RDF → Sesame repository and query with SeRQL-S. Interval join with SeRQL (including SPARQL equivalent).

Overview case studies

- We will look at two examples. Several issues that will pass the revue:
  - Reusing legacy systems and reusing SW technologies
  - Scalability
  - System interoperability and integration
  - Methodology
  - Promises, experiments, but few concrete success stories
  - Adding more semantics to non-SW applications and/or broadening SW technologies
- There are more projects

The Foundational Model of Anatomy

- Ontology of parts of the human body, developed over 6 years in Protégé, about 72000 concepts (universals) and 1.9mln relations between them [RM03].
- The ingredients:
  - A "Protégé database" in MySQL with Protégé application interface and web interface [6], "frame-based through Protégé application"
  - A version in PostgreSQL, queries with StruQL through OQAFMA [MBR03]
- A few aims: checking consistency, satisfiability, querying to find information, use it as reference ontology to use for domain ontology & software development

The FMA in the Semantic Web

How can we make the FMA "SW compliant"?

- What an RDF triple store with SPARQL can do for us
- Transforming the frames/database into an OWL-DL representation [ZBG06]
- What about having *n* versions of the FMA?
  - A well-designed RDBMS for querying
  - Triple store for people who insist on it
  - OWL 1.1 full for comprehensiveness (in modules)
  - "Lite" version with DL language of lower complexity for automated reasoning
  - Part of the FMA in ER/UML/ORM for semantically enhanced applications

The FMA in the Semantic Web

How can we make the FMA "SW compliant"?

- What an RDF triple store with SPARQL can do for us
- Transforming the frames/database into an OWL-DL representation [ZBG06]
- What about having *n* versions of the FMA?
  - A well-designed RDBMS for querying **x**
  - Triple store for people who insist on it ✓
  - OWL 1.1 full for comprehensiveness (in modules) ✓
  - "Lite" version with DL language of lower complexity for automated reasoning ✓
  - Part of the FMA in ER/UML/ORM for semantically enhanced applications **x**

The FMA in the Semantic Web

How can we make the FMA "SW compliant"?

- What an RDF triple store with SPARQL can do for us
- Transforming the frames/database into an OWL-DL representation [ZBG06]
- What about having *n* versions of the FMA?
  - A well-designed RDBMS for querying ✗ ✓ ⇒ InstanceStore [BHT05], QuOnto [Aetal05] [11]
  - Triple store for people who insist on it ✓
  - OWL 1.1 full for comprehensiveness (in modules) ✓
  - "Lite" version with DL language of lower complexity for automated reasoning ✓
  - Part of the FMA in ER/UML/ORM for semantically enhanced applications ✗ ✓ ⇒ the conceptual modelling languages are mapped into DL anyway

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

Overview 23-author article [Retal07]

- "A significant barrier to translational research is the lack of uniformly structured data across related biomedical domains."
- HCLSIG was launched to explore the application of SW technologies in a variety of areas
- "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."

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

Overview 23-author article [Retal07]

- ⇒ "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."

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 [13], 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?

More case studies and projects

- Debugging the Bug [4]: Developing complete and consistent models of metabolism for bacterial organisms. (integration, ontologies)
- FungalWeb [7]: Ontology, the Semantic Web, Intelligent Systems for Fungal Genomics. (ontologies, querying, NLP)
- HistOn [12] (RDF, Sesame, OWL)
- Cell cycle ontology [2]: knowledge integration framework
- and more

A few more SWLS topics

- **Automated reasoning**, for many different purposes [KRM07]
  - A successful case study finding new information, and what it takes to get there [WSH07]
  - "Nonstandard" use of reasoning services [BM06]
- SWLS services (e.g. [L05] [Letal04])
- SWLS and the Grid (e.g. [RJS05])
- SWLS and workflows/workbenches
- SWLS and visualization of ontologies and other knowledge
- SWLS and NLP
- Personalization [MPRB06], the "resourceome" [CMA05], or: how to organise the ('your') resources and find what you need
- The human dimension, critical mass [GW06]

Gaps (from SWLS view), solutions are work in progress

- Taking into account fuzzy, uncertainty, probability, and provenance (evidence codes) in the automated reasoning
- Methodologies for ontology development & maintenance
- Linking to data
- Better results for mining information from literature
- Distributed ontologies & querying, modularization, granularity

Summary

- More problems than solutions
- From legacy systems to SW Technologies
- Linking and integrating software systems
- SW technologies are a means (with the benefit of the doubt)
- Goal-driven: in the end, the life scientists, biomedical researchers, and healthcare practitioners need clearly demonstrated benefits that make it worthwhile the efforts (e.g., finding novel knowledge about nature, better healthcare services)

References (and references therein)

[Aea05] Acciarri, A., Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Palmieri, M., Rosati, R. 2005. QuOnto: Querying ontologies. In *Proc. of the 20th Nat. Conf. on Artificial Intelligence (AAAI 2005)*, 1670–1671.

[BM06] Bandini, S., Mosca, A. Mereological knowledge representation for the chemical formulation. *2nd Workshop on Formal Ontologies Meets Industry 2006 (FOMI2006)*, 14-15 December 2006, Trento, Italy. 55-69.

[BHT05] Bechhofer, S., Horrocks, I., Turi, D. 2005. The OWL Instance Store: System description. In *Proceedings of CADE-20*, LNCS. Springer Verlag.

[Cea07] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *J. of Automated Reasoning*, 2007. (to appear).

[CMA05] Cannata, N., Merelli, E., Altman, R.B. Time to Organize the Bioinformatics Resourceome. *PLoS Computational Biology*, 2005, 1(7): e76.

[GW06] Good, B.M., Wilkinson, M.D. The Life Sciences Semantic Web is Full of Creeps! *Briefings in Bioinformatics*, 2006, 7(3):275-286.

[HKS06] Horrocks I, Kutz O, Sattler U. The Even More Irresistible $\mathcal{SROIQ}$. In: *Proc. of the 10th International Conference of Knowledge Representation and Reasoning (KR-2006)*, UK, 2006.

[K05a] Keet, C.M. Current characteristics and historical perspective of Computer Science and IT with/for biology. *CSBio Reader: extended abstracts of the 'CS&IT with/for biology' Seminar Series*, Free University of Bozen-Bolzano, **1**, 1-8. 2005.

References (and references therein)

[K05b] Keet, C.M. Factors affecting ontology development in ecology. *Data Integration in the Life Sciences 2005 (DILS2005)*, Ludäscher, B, Raschid, L. (eds.). San Diego, USA, 20-22 July 2005. Berlin: Springer-Verlag, Lecture Notes in Bioinformatics LNBI 3615, 46-62.

[KR07] Keet, C.M., Rodriguez, M. Toward using biomedical ontologies: trade-offs between ontology languages. *AAAI 2007 Workshop on Semantic e-Science (SeS07)*, 23 July 2007, Vancouver, Canada.

[KRM07] Keet, C.M., Roos, M., Marshall, M.S. A survey of requirements for automated reasoning services for bio-ontologies in OWL. *Third international Workshop OWL: Experiences and Directions (OWLED 2007)*, 6-7 June 2007, Innsbruck, Austria.

[L05] Lefranc, M.-P. A standardized and dynamic approach for immunogenetics and immunoinformatics: IMGT-Choreography based on the IMGT-ONTOLOGY concepts. *CSBio Reader: extended abstracts of the 'CS&IT with/for biology' Seminar Series*, Free University of Bozen-Bolzano, **1**, 26-38.

[Letal04] Lord, P. Bechhofer, S., Wilkinson, M.D., Schiltz, G., Gessler, D., Hull, D., Goble, C., Stein, L. Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt. *ISWC 2004*, LNCS 3298, 350-364.

[MPRB06] Marshall, M.S., Post, L., Roos, M., Breit, T.M. Using semantic web tools to integrate experimental measurement data on our own terms. *Workshop on Knowledge Systems in Bioinformatics (KSinBIT'06)*, Springer Verlag LNCS 4277, 2006, 679-688.

[MBR03] Mork, P., Brinkley, J.F., Rosse, C. OQAFMA Query Agent for the Foundational Model of Anatomy: a prototype for providing flexible and efficient access to large semantic networks. *Journal of Biomedical Informatics*, 2003, 36: 501-517.

References (and references therein)

[RM03] Rosse, C., Mejino, J.L.V.: A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 2003, **36**: 478-500.

De Roure, D., Jennings, N.R., Shadbolt, N.R. The semantic grid: past, present, and future. *Proceedings of the IEEE*, 93(3): 669-681.

[RLetal07] 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.

[Retal07] Ruttenberg, A. et al. Advancing translational research with the Semantic Web. *BMC Bioinformatics*, 2007, 8(Suppl 3):S2.

[Setal05] Smith, B., Ceusters, W., Klagges, et al. Relations in biomedical ontologies. *Genome Biology*, 2005, 6:R46.

[WSH07] Wolstencroft, K., Stevens, R., Haarslev, V. Applying OWL reasoning to genomic data. In: *Semantic Web: revolutionizing knowledge discovery in the life sciences*, Baker, C.J.O., Cheung, H. (eds), Springer: New York, 2007, 225-248.

[ZBG06] Zhang S, Bodenreider O, Golbreich C. Experience in reasoning with the Foundational Model of Anatomy in OWL-DL. In: *Pacific Symposium on Biocomputing 2006*, Altman RB, Dunker AK, Hunter L, Murray TA, Klein TE, (Eds.). World Scientific, 2006, 200-211.

References (and references therein)

[1] Basic Formal Ontology. http://www.ifomis.uni-saarland.de/bfo/.

[2] Cell Cycle Ontology. http://www.cellcycleontology.org.

[3] D2RQ. http://sites.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/.

[4] Debugging the Bug. http://bio.freelogy.org/wiki/Debugging_the_bug.

[5] Descriptive Ontology for Linguistic and Cognitive Engineering. http://www.loa-cnr.it/DOLCE.html.

[6] Foundational Model of Anatomy (FMA). http://fme.biostr.washington.edu:8089/FME/index.html.

[7] FungalWeb. http://www.cs.concordia.ca/FungalWeb/.

[8] General Formal Ontology. http://www.onto-med.de/en/theories/gfo/index.html.

[9] OBO Foundry. http://obofoundry.org.

[10] OWL 1.1 Tractable Fragments. http://webont.org/owl/1.1/tractable.html.

[11] QuOnto. http://www.dis.uniroma1.it/∼quonto/.

[12] Semantic Data Integration for Histone Use Case Website. http://integrativebioinformatics.nl/semanticdataintegration.html.

[13] Time Ontology in OWL. http://www.w3.org/TR/owl-time/.