Carleton UNIVERSITY

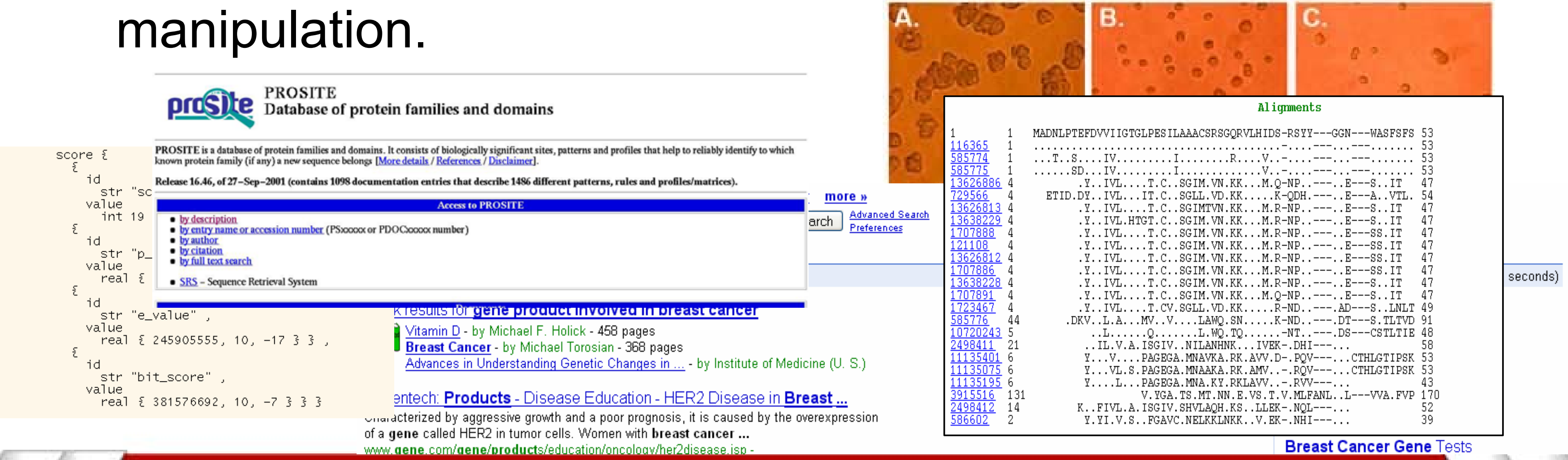
Semantic Web: from theory to practice

Natalia Villanueva-Rosales

PhD Candidate
School of Computer Science
Carleton University
nvillanu@scs.carleton.ca

Motivation

- Huge amount of information in web resources and distributed information systems.
- Heterogeneous formats, platforms, languages, identifiers.
- Data is not machine understandable, syntactic manipulation.



PROSITE
Database of protein families and domains

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs [More details / References / Disclaimer].

Release 16.46, of 27–Sep–2001 (contains 1098 documentation entries that describe 1486 different patterns, rules and profiles/matrices).

Access to PROSITE

- by description
- by entry name or accession number (PSxxxxx or PDOCxxxxx number)
- by author
- by citation
- by full text search
- SRS – Sequence Retrieval System

```
score {
  {
    id
      str "sc
    value
      int 19
  {
    id
      str "p_
    value
      real {
  {
    id
      str "e_value" ,
    value
      real { 245905555, 10, -17 } } ,
  {
    id
      str "bit_score" ,
    value
      real { 381576692, 10, -7 } } }
```

k results for **gene product involved in breast cancer**

Vitamin D - by Michael F. Holick - 458 pages
Breast Cancer - by Michael Torosian - 368 pages
Advances in Understanding Genetic Changes in ... - by Institute of Medicine (U. S.)

entech: **Products** - Disease Education - HER2 Disease in **Breast** ...
Characterized by aggressive growth and a poor prognosis, it is caused by the overexpression of a **gene** called HER2 in tumor cells. Women with **breast cancer** ...
www.**gene**.com/**gene**/**product**s/education/oncology/her2disease.jsp -

Alignments

```
1         1    MADNLPTEFDVVIIGTGLPESILAAACSRSGQRVLHIDS-RSYY---GGN---WASFSFS 53
116365    1    .......................................-....---...---....... 53
585774    1    ...T..S....IV.........I........R....V..-....---...---....... 53
585775    1    ......SD...IV.........I.............V..-....---...---....... 53
13626886  4           .Y..IVL....T.C..SGIM.VN.KK...M.Q-NP..---..E---S..IT   47
729566    4      ETID.DY..IVL...IT.C..SGLL.VD.KK.....K-QDH.---..E---A..VTL. 54
13626813  4           .Y..IVL....T.C..SGIMTVN.KK...M.R-NP..---..E---S..IT   47
13638229  4           .Y..IVL.HTGT.C..SGIM.VN.KK...M.R-NP..---..E---S..IT   47
1707888   4           .Y..IVL....T.C..SGIM.VN.KK...M.R-NP..---..E---SS.IT   47
121108    4           .Y..IVL....T.C..SGIM.VN.KK...M.R-NP..---..E---SS.IT   47
13626812  4           .Y..IVL....T.C..SGIM.VN.KK...M.R-NP..---..E---SS.IT   47
1707886   4           .Y..IVL....T.C..SGIM.VN.KK...M.R-NP..---..E---SS.IT   47
13638228  4           .Y..IVL....T.C..SGIM.VN.KK...M.R-NP..---..E---S..IT   47
1707891   4           .Y..IVL....T.C..SGIM.VN.KK...M.Q-NP..---..E---S..IT   47
1723467   4           .Y..IVL....T.CV.SGLL.VD.KK.....R-ND..---.AD---S..LNLT 49
585776    44        .DKV..L.A...MV..V....LAWQ.SN.....K-ND..---.DT---S.TLTVD 91
10720243  5             ..L......Q.......L.WQ.TQ......-NT..---.DS---CSTLTIE 48
2498411   21           ..IL.V.A.ISGIV..NILANHNK...IVEK-.DHI---...           58
11135401  6            Y...V....PAGEGA.MNAVKA.RK.AVV.D-.PQV---...CTHLGTIPSK 53
11135075  6            Y...VL.S.PAGEGA.MNAAKA.RK.AMV..-.RQV---...CTHLGTIPSK 53
11135195  6            Y....L...PAGEGA.MNA.KY.RKLAVV..-.RVV---...           43
3915516   131                   V.YGA.TS.MT.NN.E.VS.T.V.MLFANL..L---VVA.FVP 170
2498412   14          K..FIVL.A.ISGIV.SHVLAQH.KS..LLEK-.NQL---...           52
586602    2            Y.YI.V.S..FGAVC.NELKKLNKK..V.EK-.NHI---...           39
```

Breast Cancer Gene Tests

Motivation

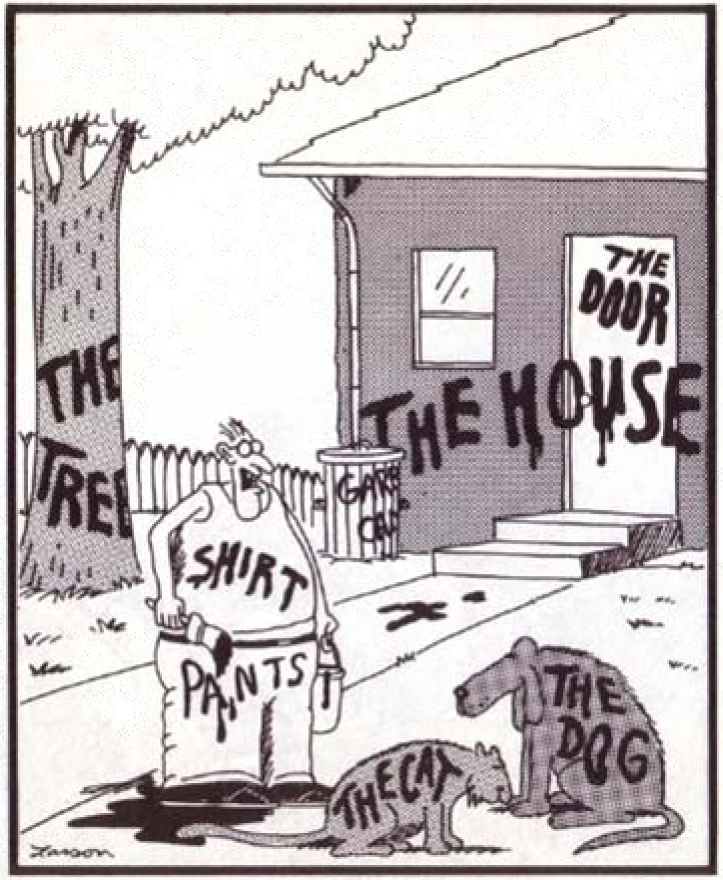
Find/answer questions on the web

- Use of background knowledge
- Access and integration of relevant information on the web.

E.g. Can a yeast scientist find *genes with transferase activity that are part of a complex and participate in a non-viable experiment?*

Semantic Web (SW)

- Extension of the current www
- Machine understandable annotations using logic-based knowledge representation languages
- Semantic agreement defined by ontologies
- Automated reasoning to obtain non-obvious inferences during information retrieval



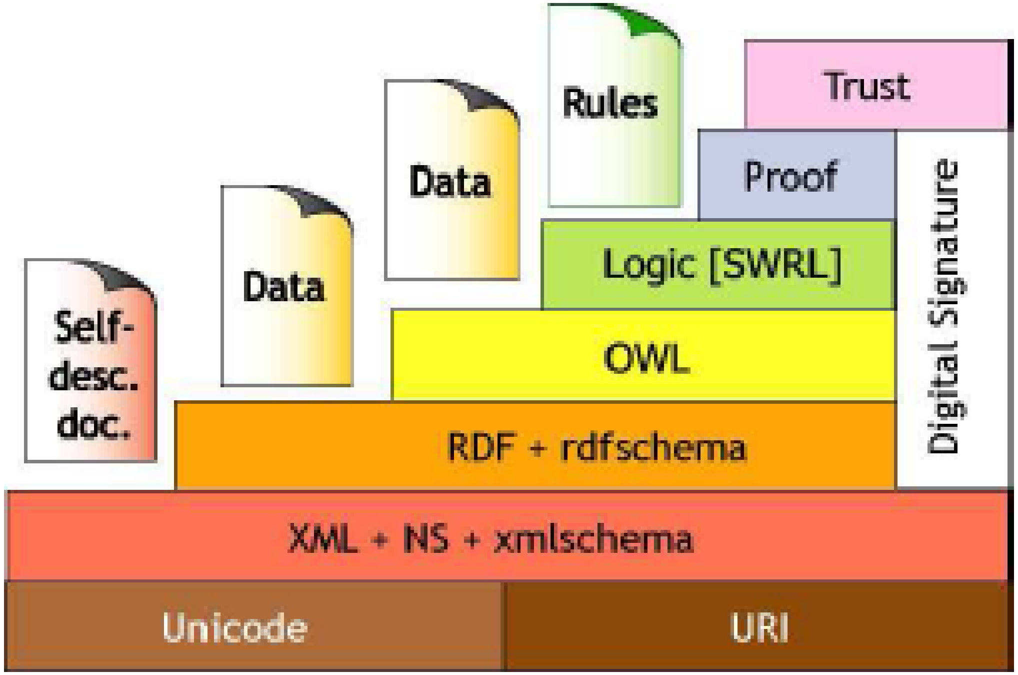
"Now! ... That should clear up a few things around here!"

Source: http://dit.unitn.it/~aiellom/images/semanticWeb.jpg

W3C Semantic Web

SW Languages

Standard markup languages



Source: Modified from Semantic Web talk by Tim Berners-Lee at XML 2000

eXchange Markup Language (XML)

- XML documents contain elements.
- Elements must have an opening and closing tag.
- Elements can have child elements (ordered) and attributes (not ordered).
- Elements must be properly nested and can be extensible.
- Example:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<message dateSent="03/01/06">
    <sender>Jack<sender>
    <recipient>Mary</recipient>
    <content>What is all this talk about?
    </content>
</message>
```

XML document syntax - 2

```
<?xml version="1.0" encoding="ISO-8859-1"?>
     <message dateSent="03/01/06">
     <title>Question</title>
     <sender>
         <title>Mr<title> <name>Jack</name>
     </sender>
    <recipient>
         <title>Miss</title> <name>Julie</name>
    </recipient>
    <content>What is all this talk about?</content>
    </message>
```

- XML documents can contain elements from different sources, it is widely used for data exchange and data integration.
- To avoid ambiguity and collision of element names name spaces are used. E.g. The first "title" refers to the message title, the second and thethird to the sender and recipient title respectively.

XML Namespaces

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<message xmlns:person="www.dumontierlab.com/Person"
  dateSent="03/01/06">
     <title>Question</title>
     <sender>
         <person:title>Mr<person:title> <name>Jack</name>
     </sender>
    <recipient>
         <person:title>Miss</person:title> <name>Julie</name>
    </recipient>
    <content>What is all this talk about?</content>
    </message>
```

- Use of namespaces as prefix, e.g. person.
- Namespace attribute is placed in the start tag of an element.
- Qualified names have scopes beyond their containing document.
- Group elements and attribute definitions from different contexts.

More XML

- XML structure validation: XML Schema.
  - Successor of DTD.
  - Elements and attributes of a document.
  - Order, number and type of child elements.
  - Data types.
- Search: XPath.
- Query: XQuery.

xmlns:xs="http://www.w3.org/2001/XMLSchema"

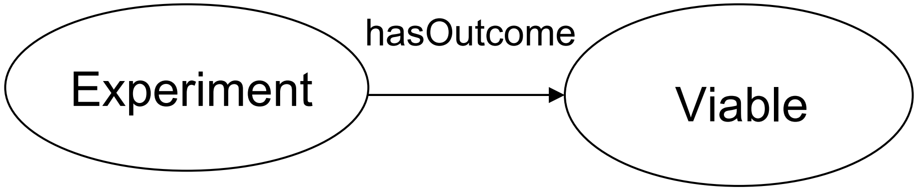
Resource Description Framework (RDF)

- Creation of data models in terms of resources and relations between them: statements.
- Basic elements in RDF:
  - Resources:
    - Anything represented by an URI (a person, a painting, an e-store).
    - Nodes in a graph representation.
    - Example: http://ontology.dumontierlab.com/Experiment
  - Properties:
    - Also represented by an URI.
    - Edges in a graph representation.
    - Binary relations between two resources.
    - http://ontology.dumontierlab.com/hasOutcome
  - Literals
    - Concrete data values.
    - Nodes in a graph representation.
    - Can use XML Schema datatypes
    - Example: "2", "blue", "10-12-2006"

RDF Syntax

RDF databases can be seen as graph databases of triples

- Graphical representation:
  - Experiment is the subject
  - hasOutcome is the property
  - and Viable is the object



- Triple representation:

(Experiment, hasOutcome, Viable)

RDF Syntax - 2

- The RDF/XML sytax:

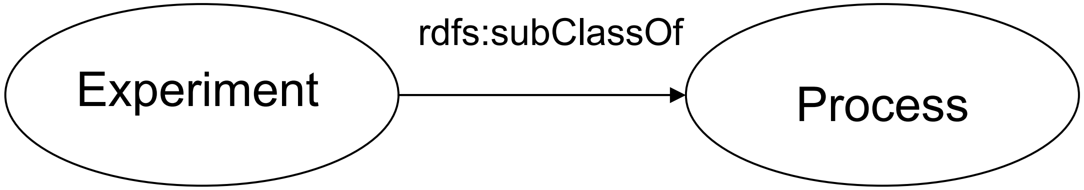
```
 <yowl:Experiment rdf:about="urn:lsid:yeastgenome.org:#experiment_1">

   <rdfs:comment
        rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        YDR163W - Exhibits sensitivity at 5 yowl:Generations when   grown in
        synthetic complete - thr medium.. Results from large scale deletion
        study. Phenotype determined by fitness score.
   </rdfs:comment>
   …
   <yowl:hasOutcome rdf:resource="urn:lsid:yeastgenome.org:#phenotype_1"/>
</yowl:Experiment>
```

http://ontology.dumontierlab.com/yowl-hcls

RDF Schema

- Classes and instances of classes.
- Class hierarchy: subclass of
  - Properties of a class are inherited by instances that to a subclass.
- Property hierarchy: subproperty of.

- Property domain and range.



Experiment

rdfs:subClassOf

Process

http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

Triple stores

- Triples can be collected together into a triple store, e.g.:
  - Sesame
    http://www.openrdf.org/
  - Jena
    http://jena.sourceforge.net/DB/layout.html
  - 3Store
    http://sourceforge.net/projects/threestore/

- Query languages:
  - SPARQL (W3C recommendation)
    http://www.w3.org/TR/rdf-sparql-query/
  - SeRQL
    www.openrdf.org/doc/**SeRQL**manual.html
  - RQL
    http://139.91.183.30:9090/RDF/RQL/
  - …

- Nice scalability.

RDF

- RDF + RDFS is already an ontology language.
- Some nice features: reification.
- Used for annotations.
- Lack of negation (disjunction, cardinality constraints)
- First level above RDF:
  - Ontology language to formally describe the meaning of terminology used in Web resources (and real world)
  - The Web Ontology Language (OWL) provides greatermachine interpretability of Web content than the one supported byXML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formally specified semantics to describe ontologies.

Web Ontology Language (OWL)

- An ontology is a “(shared) specification of a (knowledge) conceptualization‟. Gruber 1993.

- An ontology contains statements like “Enzime is a Protein that catalizes some chemical reaction”.

- Reuses the semantics of some RDF elements (i.e. subsumption, domain and range) with a richer set of primitives:
  - Relations between classes: equality, enumerated classes, disjointness. E.g. An enzyme is not a reaction.
  - Cardinality restrictions axioms, e.g. A human hand has five fingers
  - Richer characteristics of properties (e.g. symmetry, transitivity, reflexivity)

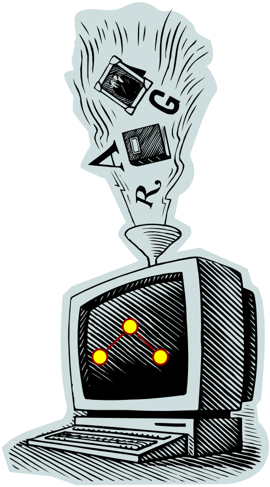
http://www.w3.org/2002/07/owl

Three sublanguages of OWL

- Lite: Supports classification hierarchy and simple constraints (restricted cardinality constraints).
- DL: Based on Description Logics, decidable fragment of First Order Logic (FOL)
  - Balance between decidability and expressivity
  - Sound and complete algorithms implemented for reasoning problems.
  - Complex concepts are defined out of simpler ones.
- Full: Allows the use OWL language constructs with no restrictions, undecidable.

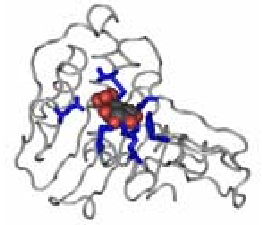
http://www.w3.org/2002/07/owl

OWL ontologies in use



Semantic Query Answering over Statistical Graphs

Leo Ferres [1], Michel Dumontier [2,3], **Natalia Villanueva-Rosales** [3]
[1]Human-Oriented Technology Laboratory, [2]Department of Biology,
[3]School of Computer Science, Carleton University



Dumontier Lab

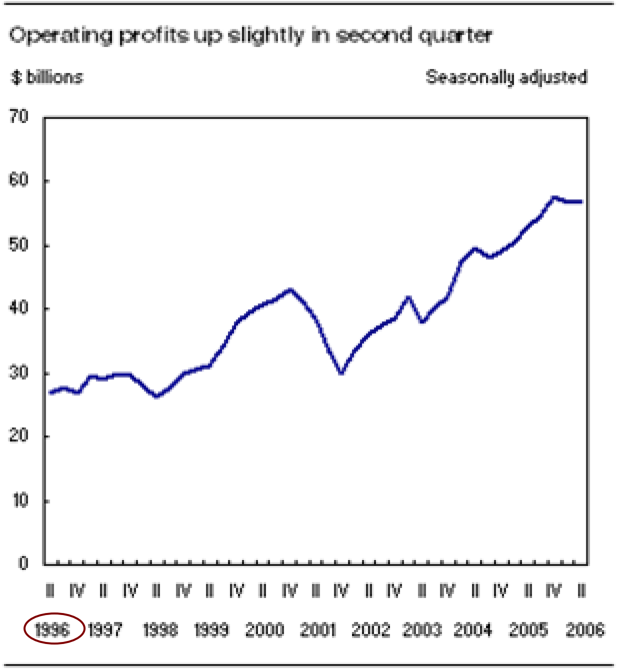
Supported by: Statistics Canada, Cognos Inc., NSERC.

Motivation

Statistical Graphs

- Used for data visualization in most enterprises.
- Created with a communicative purpose.
- Stored as unstructured images or binary objects (lack of expressive semantics).
  - Limited exchange or integration (merging) of complementary graphs
  - Retrieve graphs with particular content and
  - Question answering across graphs.
  - Not easily accessible in certain contexts (small devices, visually impaired people)

Motivation



What does "1996" mean?
Integer, string, …

```
...
 <category id="0">
 II
 <secondary>1996</secondary>
 </category>
 <value id="0">26.7</value>

 ...
```

In XML/XMLS?

Datatypes, restricted values, arbitrary name elements, no meaning.

gSem Objectives

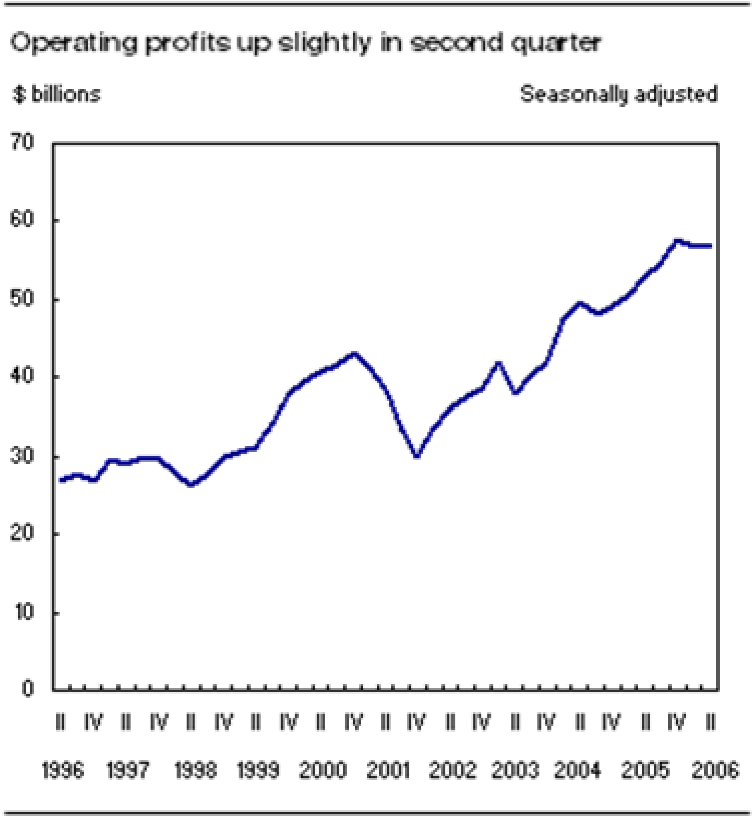
1. Improve the efficiency of searching statistical graph knowledge
2. Facilitate sophisticated question answering about statistical graph knowledge
3. Increase the accessibility of statistical graphs for the visually impaired, and
4. Enable graph re-purposing beyond their original communicative intent.

This solution

OWL ontology for representation, integration, exchange, and semantic query answering of statistical graph data using automated reasoning.

?

This solution



A Graph has a Primary title and a Plot.
A Plot has X and Y axis.
X axis has a Primary category and a Category data axis.
X axis might have a Secondary Category.
A Plot has Series.
Series have at least 2 Data points.
Data points have category data and value data.
…

Manchester OWL syntax [Horridge et al., OWLED 06].

Methodology

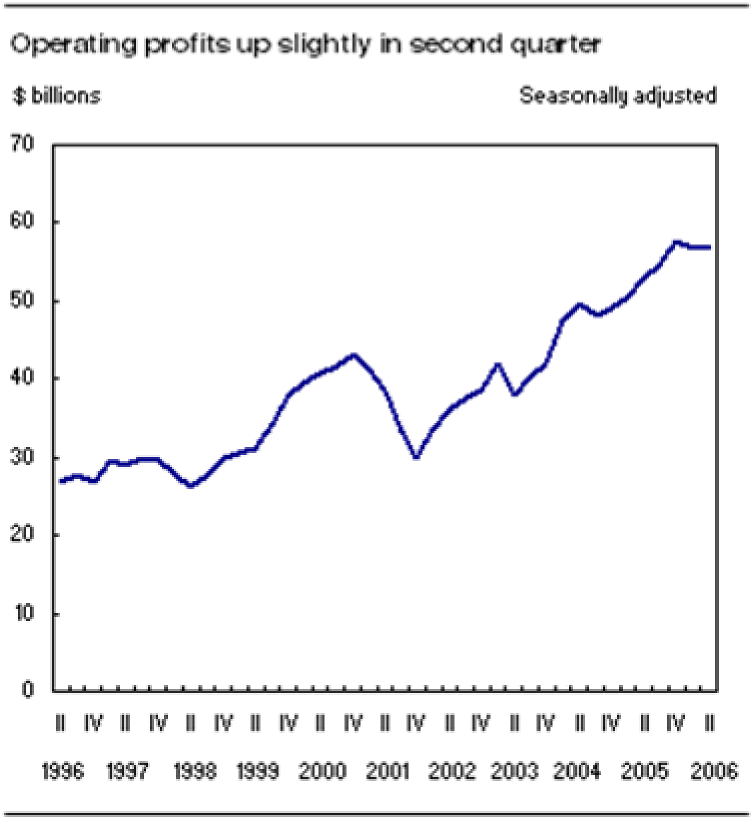
Requirements

- iGraph-Lite (NLP interface to an enriched XML representation) [Ferres et. al., ICCHP 06]
- Competency questions [Uschold et. al., KE 96]
  - What is the graph title?
  - Do the x-axis categories of a graph correspond to years?
  - Which quarters between 2006 and 2007 have increasing sales?
  - Which graphs contain information of sales, excluding serving sales?

Methodology - 2

Corpus of graphs

- Statistics Canada “The Daily”
- 10 years (1996 – 2006)
- 5060 graphs
  - 2883 line Graphs (63.3%)



Operating profits up slightly in second quarter

Methodology - 3

- Deconstruction of graphs, collection of concepts.
- Added concepts in the OOXML, ODF and *GraphRep* ontology (coverage).
- Creation of simple taxonomy (is-a relation).
- Definition of relations between objects
- Mapped to upper level ontologies (*BFO, BRO*).
- Three layer modelling approach.
- OWL 1.1 specification, *SHOIQ (D)* .

http://www.webont.org/owl/1.1/xml_syntax.html#ref-owl-1.1-specification

Layer 1: Statistical Graph Ontology (SGO)

- Describes a graph (its components and how they are related).
- 62 classes (concepts)
  - E.g. Graph,Title, CategoryData, etc.
- 22 properties
  - 16 object properties. E.g. hasTitle.
  - 6 datatype properties. Eg. hasValue.
- Definitions of the form:

  A Graph hasPart one or more Plot and may hasTitle one or more Title and hasSource Source of origin.

http://ontology.dumontierlab.com/statistical-graph-primitive

Layer 2: Augmenting the SGO

- More complex definitions using more restrictive operations like disjunction, union, intersection, class equivalence, existential and universal restrictions and qualified cardinality restrictions (inferences using classification and realization).

  GraphTitle is equivalent to a Title that isTitleOf some Graph

- Contextual Knowledge
  - Ontology mapping to upper level ontologies (BFO and BRO) and time interval ontology. Ontology reuse.

    E.g. FirstQuarter, Quarter, Year.

- Restrictions hold in any application.

http://ontology.dumontierlab.com/statistical-graph-complex
http://ontology.dumontierlab.com/time-interval

Layer 3: Augmenting the SGO - 2

iGraph Requirements

- Application dependent restrictions, not expected to hold outside the application.
  - E.g. LineGraph

- Useful for data exchange.

http://ontology.dumontierlab.com/statistical-graph-igraph

Ontology Population

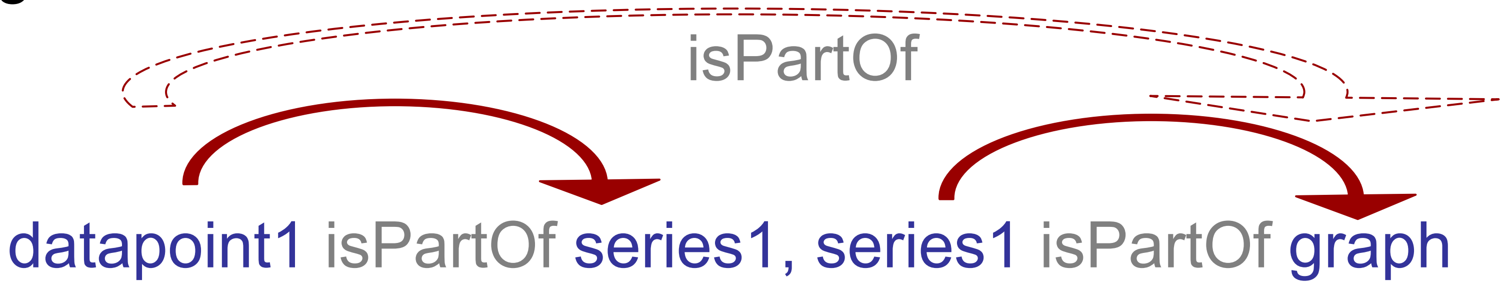
- From iGraph-Lite
  - Enriched XML documents after being annotated by iGraph-Lite with x-axis categories, titles, etc.
- To concept instances in the ontology using OWL RDF/XML syntax. E.g. “1997” is an instance of Year.
- XLST transformations.

Semantic QA over Time-Series Graphs

- Q1: Retrieve all the datapoints in the graph.

DataPoint that isPartOf some Graph

E.g.



datapoint1 isPartOf series1, series1 isPartOf graph

Transitivity.

Using Protégé 4 alpha (build 53) , FACT++ DL reasoner and Manchester Syntax.

Semantic QA over Time-Series Graphs

- Q2 : Retrieve all the value data for the second quarter of any year.

ValueData that isPartOf some
(DataPoint that hasPart some (SecondQuarter))

E.g.

y7 isPartOf datapoint7, datapoint7 hasPart x7,
x7 type SecondQuarter

Ontology mapping.

Using Protégé 4 alpha (build 53) , FACT++ DL reasoner
and Manchester Syntax.

Semantic QA over Time-Series Graphs - 2

- Q3 : Retrieve all series that contain time intervals (time series).

TimeSeries ≡ Series that hasPart some TimeInterval

E.g.

series1 hasPart datapoint7, datapoint7 hasPart x7,
x7 type SecondQuarter,
SecondQuarter subClassOf Quarter,
Quarter subClassOf TimeInterval

Class def.

Using Protégé 4 alpha (build 53) , FACT++ DL reasoner and Manchester Syntax.

Semantic QA over Time-Series Graphs (2)

- Q4: Retrieve all time series graphs.

TimeSeriesGraph ≡ Graph that hasPart some TimeSeries

E.g.

series1 hasPart datapoint7, datapoint7 hasPart x7,
x7 type SecondQuarter,
SecondQuarter subClassOf Quarter,
Quarter subClassOf TimeInterval,
graph hasPart series1

Across graphs.

Using Protégé 4 alpha (build 53) , FACT++ DL reasoner and Manchester Syntax.

Other implications

Data exchange

- Standard (syntactic) representation of the data for machine consumption.
- Unambiguous meaning (semantics).
- Three layer flexible model.
- Exchange between statistical agencies.
- XSLT to return new information to applications.

Ongoing work

- Provide semantic query answering over time-series graphs. [Ferres et. al., VORTE07]
- Data model for graph information integration and exchange.
- Increased explicit knowledge in statistical graphs for certain demographics (accessibility).
- Enhanced iGraph-Lite using Semantic Web.
- Still, some challenges:
  - Spatial and temporal reasoning.
  - "Type" of graphs according to their content [Dumontier et. al, Submitted].

Modeling the Pharmacogenomics of Depression

Michel Dumontier[1], Muhammad Faizan[2], Joseph Obeng[1],

Natalia Villanueva-Rosales[2]

[1]Department of Biology, [2]School of Computer Science
Carleton University

Motivation

Pharmacogenomics

- Pharmacological response of a drug with respect to genetic variation .
- Answer questions about therapeutic, pharmacological or genetic aspects
- Essential to the delivery of better health care
- Towards personalized medicine: provide the most effective therapeutic strategy based on physiological and genetic factors.

This solution

Reasoning capable knowledge base, with OWL-DL ontologies that capture the knowledge accumulated by PharmGKB [Hernandez-Boussard, NAR, 2008] for sophisticated query answering. over that knowledge.

?

Requirements

- PharmGKB database [Ferres et. al., ICCHP 06]
  - Genes
  - Gene variants
  - SNPs
  - Drugs
  - Measures and outcomes
  - Gene-drug interactions
  - Drug treatments.

- Augmentation with literature curated pharmacogenomics knowledge of depression.

- Competency questions [Uschold et. al., KE 96]
  - What is the most effective drug treatment for an individual with a particular genetic profile that suffers from a particular disease?
    - Which drugs yield a favorable outcome?
    - What are the possible side effects?Do the x-axis categories of a graph correspond to years?
  - Which gene variants affect therapeutic outcomes?

Methodology

- Essential concepts: e.g. Pathway, Drug.
- Ontology reuse: Biological Measure
- Creation of simple taxonomy (is-a relation).
- Definition of relations between objects
- Mapped to upper level ontologies (*BFO*, *BRO*).
- Three layer modelling approach.
- OWL 1.1 : role composition.

http://ontology.dumontierlab.com/biological-measure-primitive

Layer 1: Pharmacogenomics Primitive

- 40 classes (concepts)
  - E.g. DrugTreatment, BiologicalMeasure, Drug, etc.
- Domain dependent properties
  - Eg. hasVariant.

http://ontology.dumontierlab.com/pharmacogenomics-primitive

Layer 2: Pharmacogenomics Complex

- Complex class definitions

  DrugGeneInteraction is equivalent to a Process that hasParticipant some Drug and hasParticipant some Gene and not DrugTreatment

- Role chains

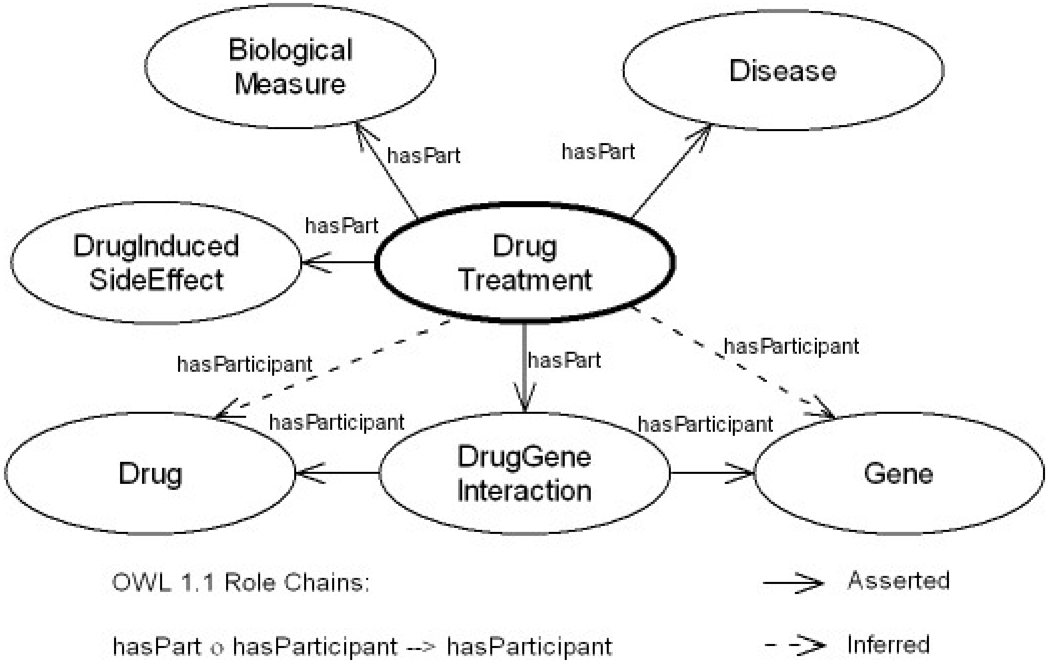
  hasPart o hasParticipant --> hasParticipant

- Ontology mapping to upper level ontologies (BFO and BRO), biological measure and unit ontology. Ontology reuse.

  E.g. ClinicalOutcome, GenotypeMeasure, MetabolicMeasure, WeightUnit.

http://ontology.dumontierlab.com/pharmacogenomics-complex
http://ontology.dumontierlab.com/biological-measure-primitive
http://ontology.dumontierlab.com/unit-ontology-individuals

Drug Treatment



Core concepts and relations in the Pharmacogenomics Ontology

Ontology Population

- From PharmGKB web services
  - Genes, drugs, and diseases having pharmacogenomics relevance
- Pharmacogenomics of depression ontology
  - Manually curated over 40 publications,
- In-house developed java parsers.

http://ontology.dumontierlab.com/pharmacogenomics-depression

Application scenario (Semantic QA)

Treatment for an elderly patient diagnosed with depression with no postural hypothension as a side effect.

- Drugs that might lead to postural hypothension

Drug that isParticipantIn some (DrugTreatment that hasPart some (DrugInducedSideEffect that hasParticipant value PosturalHypotension and hasParticipant some SideEffectRate))

Results include Nortriptyline that exhibits a 0% side effect rate for postural hypothension.

Application scenario (Semantic QA) - 2

- Drug treatment with Notriptyline (pgkb:PA450657), knowing that our patient genotyping results indicate that our patient is homozygous at the 3435 position of the ABCB1 GENE.

DrugTreatment that hasPart some (DrugInducedSideEffect that hasParticipant value PosturalHypotension) and hasParticipant value PA450657 and hasParticipant value ABCB1_3435_C

Results include NortriptylineABCB1Treatment1, which together with more specific queries provide a recommended dose of "103 mg"

Other projects

- SMART : web-based intuitive interface for ontology-driven semantic query answering on biological knowledge.
  http://smart.dumontierlab.com
- Semi-automated ontology population
- Spatial and temporal reasoning
- Ontology mapping
- Ontologies
  http://dumontierlab.com/index.php?page=ontologies
- Visit www.dumontierlab.com

References

[Ferres et. al., ICCHP 06]	L Ferres, A Parush, S Roberts, Lindgaard. **Helping People with Visual Impairments Gain Access to Graphical Information Through Natural Language: The iGraph System**. In Proceedings of the *10th International Conference on Computers for Handicapped Persons*, LNCS, Springer.Verlag, 2006.
[Uschold et. al., KE 96]	M Uschold and M Grüninger. **Ontologies: principles, methods, and applications**. *Knowledge Engineering Review,* 1996.
[Horridge et. al., OWLED06]	M Horridge, N Drummond, J Goodwin, A Rector, R Stevens, H Wang. **The manchester owl syntax.** 2006. In Proceedings of the OWL: Experiences and Directions Workshop Series
[Dumontier et. al., WOMO07]	M Dumontier, N Villanueva-Rosales. **Three-Layer OWL Ontology Design.** 2007. *Second International Workshop on Modular Ontologies (WOMO07)*, colocated with Knowledge Capture (KCAP2007), Whistler, Canada.
[Ferres et. al., VORTE07]	L Ferres, M Dumontier, N Villanueva-Rosales. **Semantic Query Answering with Time-Series Graphs.** 2007. *The 3rd International Workshop on Vocabularies, Ontologies and Rules for The Enterprise (VORTE 2007)*, colocated with The 11th IEEE International EDOC Conference (EDOC 2007), Annapolis, USA
[Dumontier et. al, Submitted]	M Dumontier, L Ferres, N Villanueva-Rosales. **Semantic annotation and question answering of statistical graphs**. *Submitted.*

References

[Dumontier et. al., HCLS 08]	M Dumontier, M Faizan, J Obeng, N Villanueva-Rosales. **Modeling the Pharmacogenomics of Depression**. 2008. *Semantic Web for Health Care and Life Sciences* (HCLS 2008), colocated with World Wide Web Conference (WWW2008), Beijing, China.
[Batista et. al., SWC 07]	A De Leon Battista, N Villanueva-Rosales, M Palenychka, M Dumontier. **SMART: A Web-Based, Ontology-Driven, Semantic Web Query Answering Application**. 2007. *Semantic Web Challenge, International Semantic Web Conference (ISWC 2007)*, Busan, South Korea.
[Dumontier et. al., OWLEDDC 08]	M Dumontier, N Villanueva-Rosales. **Modeling Life Science Knowledge with OWL 1.1.** 2008. *OWL Experiences and Design (OWLED-DC 2008),* Washington DC, USA.

Presentation template: http://www.engsoc.org/~crans/MAE_downloads/

Thank you!

Questions?