“Shopping for data should be as easy as shopping for shoes!!”

Carole Goble
Web 2.0 + Web 3.0 = Web 5.0?

The HSFBCY + CIHR + Microsoft Research SADI and CardioSHARE Projects

Mark Wilkinson & Bruce McManus
Heart + Lung Institute
iCAPTURE Centre, St. Paul’s Hospital, UBC
SPARQL queries and “non-logical” reasoning over distributed data that doesn’t exist
Middleware
Robert Stevens
Shouldn’t be seen!
Good Middleware enables useful and exciting behaviours!
Cool!
For cardiovascular researchers
The Problem
The Problem
Align the promoters of all serine threonine kinases involved exclusively in the regulation of cell sorting during wound healing in blood vessels.

Retrieve and align 2000nt 5' from every serine/threonine kinase in Mus musculus expressed exclusively in the tunica [I | M | A] whose expression increases 5X or more within 5 hours of wounding but is not activated during the normal development of blood vessels, and is <40% homologous in the active site to kinases known to be involved in cell-cycle regulation in any other species.
The Problem
The Problem
The Solution??
Why not?
So... what can we do?
What we need
How do we make knowledge machine-readable?
Ontologies!
Ontology (Gr: “things which exist” + -logy)
An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them.
Problem…
Ontology Spectrum

Catalog/ID

Thesauri
“narrower term” relation

Terms/glossary

Informal
is-a

Formal
is-a

Formal instance

Value Restrs.

Frames
(Properties)

Selected Logical Constraints
(disjointness, inverse, …)

General Logical constraints

WHY?

Because I say so!

Because it fulfils XXX

Description in: www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-abstract.html
Clay Shirky

“Ontology is over-rated”

http://www.shirky.com/writings/ontology_overrated.html
My Definition of Ontology  
(for this talk)

Ontologies explicitly define the things that exist in “the world” based on what properties each kind of thing must have
Ontology Spectrum

- Catalog/ID
- Terms/glossary
- Thesauri
  - “narrower term” relation
- Informal is-a
- Formal is-a
- Formal instance
- Value Restrs.
- Frames (Properties)
- Selected Logical Constraints (disjointness, inverse, …)
- General Logical constraints
My goal with this talk: the “sweet spot”
COST

Selected Logical
Constraints
(disjointness,
inverse, …)

Frames
(Properties)

General
Logical
constraints

Value
Restrs.

Informal
is-a

Formal
instance

Formal
is-a

Thesauri
“narrower
term”

relation

Terms/
glossary

Catalog/
ID
COMPREHENSIBILITY

- Catalog/ID
- Terms/glossary

- Thesauri
  - "narrower term" relation

- Informal is-a

- Formal is-a

- Formal instance

- Frames (Properties)

- Value Restrs.

- Selected Logical Constraints (disjointness, inverse, …)

- General Logical constraints
Likelihood of being “right”

- Thesauri
  - “narrower term” relation
- Informal is-a
- Formal instance
- Frames (Properties)
- Value Restrs.
- Selected Logical Constraints (disjointness, inverse, ...)
- General Logical constraints

Terms/glossary

Formal is-a

Logical constraints

Catalog/ID
Here’s my argument...
An information system where machines can receive information from one source, *re-interpret* it, and *correctly use it* for a purpose that the source had not anticipated.
If we cannot achieve those two things, then IMO we don’t have a “semantic web”, we only have a distributed (??), linked database... and that isn’t particularly exciting...
Where is the Semantic Web?

REASON: “Because I say so” is not open to re-interpretation
SUNDAY WORSHIP
10:30 AM
THERE ARE SOME QUESTIONS THAT CAN NOT BE ANSWERED BY GOOGLE

ALL WELCOME

PRESBYTERIAN CHURCH
ACCESSING THE DEEP WEB

Attempting to locate and quantify material on the Web that is hidden from typical search techniques.

The Web has been rapidly “deepened” by massive databases online and current search engines do not reach most of the data on the Internet [4]. While the surface Web has linked billions of static HTML pages, a far more significant amount of information is believed to be “hidden” in the deep Web, behind the query forms of searchable databases, as Figure 1(a) conceptually illustrates. Such information may not
If we’re going to build the Semantic Web, we must make Deep Web data a priority from the outset
Based on the numerous observations we made while watching the usage and behaviour of BioMoby since 2002.
SADI Observation #1:

Most Web Services in Bioinformatics create implicit biological relationships between their inputs and associated outputs.
SADI Observation #1:

BRCA1 → hasDNASequence → Web Service → "AGCTTAGCCA..."
SADI Proposition

Making the implicit explicit allows Web Services to be a source of semantically-grounded Triples
SADI Proposition

A Web Services registry that provides discovery based on these biological relationships enables the behaviours required of the Semantic Web in Healthcare and Life Science
Dynamic
Discovery
Distributed
Interpretation
Re-interpretation
Demo #1
a plug-in to the
IO-Informatics Knowledge Explorer
The Sentient Knowledge Explorer from IO Informatics

A Semantic Integration, Visualization and Search Tool
SADI Plug-in to the Sentient Knowledge Explorer

• The Knowledge Explorer has a plug-ins API that we have utilized to pass data directly from SADI to the KE visualization and exploration system

• Data elements in KE are queried for their “type” (rdf:type); this is used to discover additional data properties by querying the SADI Web Service registry
BACKGROUND of DATASET

NIST toxicity compendium study

8 toxicants
Liver
Serum
Urine
Genomic [Affymetrix MA]
Metabolomic [LC/MS] data

• conducted under NIST Advanced Technology Program (ATP), Award # 70NANB2H3009 as a Joint Venture between Icoria / Cogenics (Division of CLDA) and IO Informatics.
• Microarray studies were conducted under NIEHS contract # N01-ES-65406.
• The Alcohol study was conducted under NIAAA contract # HHSN281200510008C in collaboration with Bowles Center for Alcohol Studies (CAS) at UNC.
• This work was also made possible through contributions from members of IO Informatics’ Working Group on “Semantic Applications for Translational Research”.

IO informatics
DEMO
Knowledge Explorer
Plug-in

For more information about the Knowledge Explorer surf to: http://io-informatics.com
DEMO
Knowledge Explorer Plug-in

For more information about the Knowledge Explorer surf to: http://io-informatics.com
Imagine there is a “virtual graph” connecting every conceivable input for every conceivable Web Service to their respective outputs

How do we query that graph?
“SHARE”

Semantic Health And Research Environment
SADI client application
Recap
what we just saw

A standard SPARQL query was entered into SHARE – a SADI-based query engine

The query was interpreted to extract the properties being queried and these were passed to SADI for Web Service discovery

SADI searched-for, found, and accessed the databases and/or analytical tools required to generate those properties
Recap
what we just saw

We posed, and answered a complex database query

WITHOUT A DATABASE

(in fact, the data didn’t even have to exist...)

The Holy Grail:

Align the promoters of all serine threonine kinases involved exclusively in the regulation of cell sorting during wound healing in blood vessels.

Retrieve and align 2000nt 5' from every serine/threonine kinase in *Mus musculus* expressed exclusively in the tunica [I | M |A] whose expression increases 5X or more within 5 hours of wounding but is not activated during the normal development of blood vessels, and is <40% homologous in the active site to kinases known to be involved in cell-cycle regulation in any other species.
RDF graph browsing and querying is useful…

…but where’s the semantics??
CARDIO
SHARE
Data + Knowledge for Cardiac Researchers
Powered by SADI

HEART & STROKE FOUNDATION OF BC & YUKON
Finding answers. For life.

Founding partner
Data exhibits “late binding”
Late binding:

“purpose and meaning” of the data is not determined until the moment it is required
Benefit of late binding

Data is amenable to constant re-interpretation
How do we achieve this?

**DO NOT**

**PRE-CLASSIFY DATA!**

just hang properties on it
These are indexing/annotation systems

These are axioms that enable classification

Selected Logical Constraints

General Logical constraints

Frames (Properties)

Value Restrs.

Formal instance

Formal is-a

Thesauri

Terms/glossary

Catalog/ID

“narrower term” relation

“disjointness, inverse, …” constraints

These are indexing/annotation systems
Design OWL-DL ontologies describing cardiovascular concepts

Ontologies are in the “Frames” area of the Ontology spectrum, and therefore can leverage SADI and be “executed” as workflows
What the…???

Did you just say “execute ontologies as workflows?!”
SADI Observation #2:

Ontology = Query = Workflow
QUERY:

Concept:

SELECT images of mutations from genes in "organism XXX" that share homology to this gene in organism YYY
As OWL Axioms

HomologousMutantImage is owl:equivalentTo {
  Gene Q hasImage image P
  Gene Q hasSequence Sequence Q
  Gene R hasSequence Sequence R
  Sequence Q similarTo Sequence R
  Gene R = “my gene of interest”
}
Class: homologous mutant images
Retrieve homologous mutant images for gene XXX
Demo #3
Discover instances of OWL classes from data that doesn’t exist...
The Holy Grail:

Align the promoters of all serine threonine kinases involved exclusively in the regulation of cell sorting during wound healing in blood vessels.

Retrieve and align 2000nt 5' from every serine/threonine kinase in *Mus musculus* expressed exclusively in the tunica [I \ M \ A] whose expression increases 5X or more within 5 hours of wounding but is not activated during the normal development of blood vessels, and is <40% homologous in the active site to kinases known to be involved in cell-cycle regulation in any other species.
SADI and CardioSHARE

OWL Class restrictions converted into workflows

SPARQL queries converted into workflows

Reasoning happening at the same time as queries are being executed
(AFAIK)

This is a completely novel behaviour!
Ontologies are reasoned
Workflows are executed
Queries are resolved

so what do we call what this client does??
“Reckoning”

Dynamic discovery of instances of OWL classes through synthesis and invocation of a Web Service workflow capable of generating data described by the OWL Class restrictions, followed by reasoning to classify the data into that ontology
But OWL Classes can be (and often are) completely “made-up”

What does that mean in the context of SADI and CardioSHARE?
Current Research

We believe that **ontologies** and **hypotheses** are actually the same “thing”

Simply assertions about individuals that may or may not exist

If an instance of a class exists, then the hypothesis is “validated”
Current Research

Constructing OWL classes that represent patient categories mimicking clinical outcomes research experiments done on the BC Cardiac Registry

The original experiments have already been published, and therefore act as a gold-standard

We attempt to use Reckoning to automatically discover instances of these patient Classes and compare our results with those of the clinical researchers
CardioSHARE architecture: Increasingly complex ontological layers organize data into richer concepts, even hypotheses.

Hypothesis
Ischemia
Hypertension
Blood Pressure

Database 1
Database 2
Analysis AlgorithmXX

SADI Web “agents”
Recap

SADI interprets queries
(SPARQL + OWL Class Definitions)

Determine which properties are available, and which need to be discovered/generated

Discovery of services via on-the-fly “classification” of local data with small OWL Classes representing service interfaces
Recap

CardioSHARE encapsulates individual data retrieval and analysis workflows into OWL Classes

Web Services expose themselves as an ontology of 1 class ("input")

Very low-cost, high accuracy

...hard to create an inconsistent ontology of one class ;-(
An information system where machines can receive information from one source, *re-interpret* it, and *correctly use it* for a purpose that the source had not anticipated.
What SADI supports

Re-interpretation:

The SADI data-store simply collects properties, and matches them up with OWL Classes in a SPARQL query and/or from individual service provider’s Web Service interface.
What SADI supports

Novel re-use:

Because we don’t pre-classify, there is no way for the provider to dictate how their data should be used. They simply add their properties into the “cloud” and those properties are used in whatever way is appropriate for me.
Important “wins”

Data remains distributed – no warehouse!

Data is not “exposed” as a SPARQL endpoint → greater provider-control over computational resources

Yet data *appears to be* a SPARQL endpoint... no modification of SPARQL or reasoner required.
Recap

CardioSHARE will allow researchers to:

- **Ask questions** in natural, intuitive ways

- **Execute complex analyses** without a bioinformatician

- **Access output from databases and analytical tools** in *exactly* the same way

- **Share hypotheses** and models in an EXPLICIT manner

- **Evaluate other’s hypotheses** over your own data
Join us!

SADI and CardioSHARE are Open-Source projects

Come join us – we’re having a lot of fun!!

http://sadiframework.org
Fin
LONGMONT, CO—The Information Age was dealt a stunning blow Monday, when a factual error was discovered on the Internet. The error was found on TedsUltimateBradyBunch.com, a *Brady Bunch* fan site that incorrectly listed the show's debut year as 1968, not 1969.

Caryn Wisniewski, a Pueblo, CO, legal secretary and diehard *Brady Bunch* fan, came across the mistake while searching for...