Decision Support for Agricultural Consultants with Semantic Data Federation and Linked Data

Mohammad Sadnan Al-Manir[1], Bruce Spencer[2,3] and Christopher J.O. Baker [1,3]

[1] Department of Computer Science, University of New Brunswick, Saint John, Canada

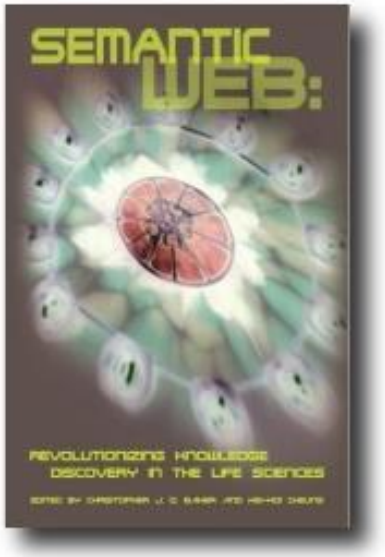
[2] Faculty of Computer Science, University of New Brunswick, Fredericton, Canada

[3] IPSNP Computing Inc, Canada

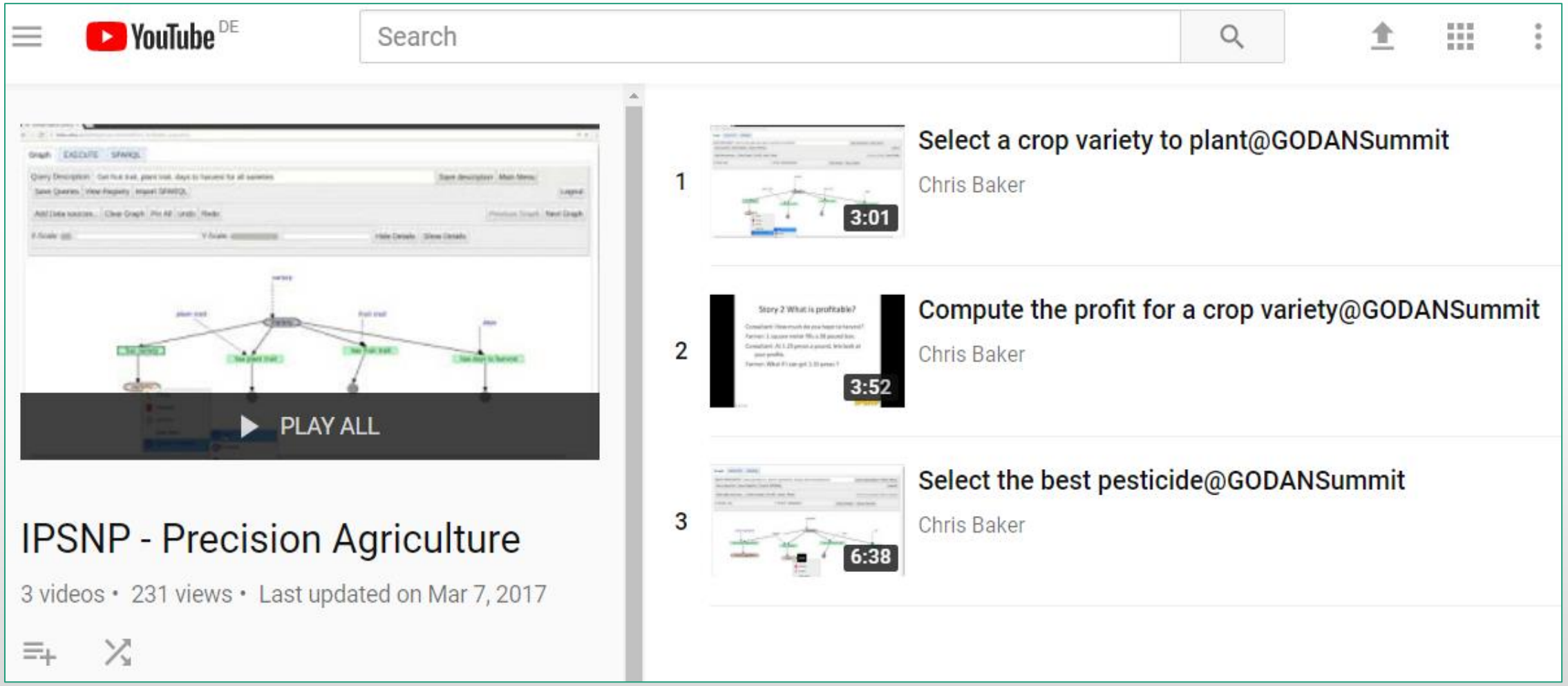
EST. 1785 UNB
UNIVERSITY OF NEW BRUNSWICK

CHRIS BAKER

- BSc Hons in Agriculture and Environmental Science
- PhD Soil Microbiology 1996
- 10 Years Wet Lab Molecular and Microbiology
- 6 years Group Leader Bioinformatics
- 2 Years Principal Investigator Data Mining Dept.
- 4 Years Research Chair – KM and Semantics
- 9 Years Professor at University of New Brunswick, CA
- *1st book on Semantic Web Technologies in Life Sciences*
- CSO of Knorex PTE Ltd., Singapore, 2009 -11
- CEO IPSNP Computing Inc. Est. 2011



SHOW THE DEMO FIRST OR LAST ?



https://tinyurl.com/IPSNP-HYDRA-Videos-Agric

End Users and End Use Challenges

- Policy makers and prog. managers rely on **data at multiples scales**.
- Have mission **critical tasks** with bespoke needs on an **ad-hoc basis**.
- Data not accessible in one location.
- Unable to compose complex queries e.g. SPARQL syntax
- **Needs help from a data scientist !**

- Findable / Discoverable – poorly described / indexed data repositories.
- Data is not OPENly **A**ccessible.
- **I**nteroperable: Diverse formats, nomenclatures, lack of semantics.
- **R**eusable: Fit for purpose ? Licensing ?
- Distributed Data Silos.
- Unstructured / raw text data.

AN AGRICULTURE POLICY MAKER WANTS

ALERTS

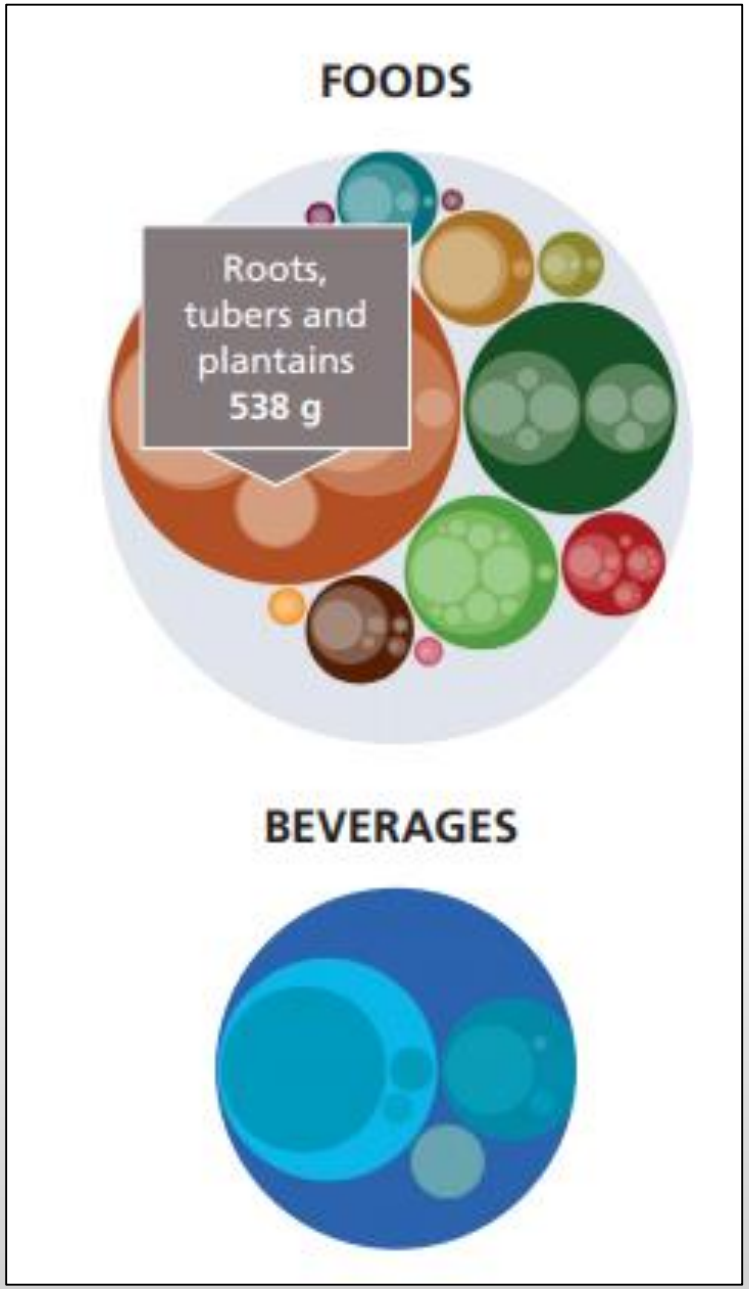
- A Ministry of Agriculture wants to promote a bio-fortified variety of a staple crop shown to be very productive in smallholder settings

REACTION

- Need to know the current level of consumption of the conventional varieties of this crop in the different age and gender.

DATA MINING

- Generate age- and sex- disaggregated data on food consumption of staple crops



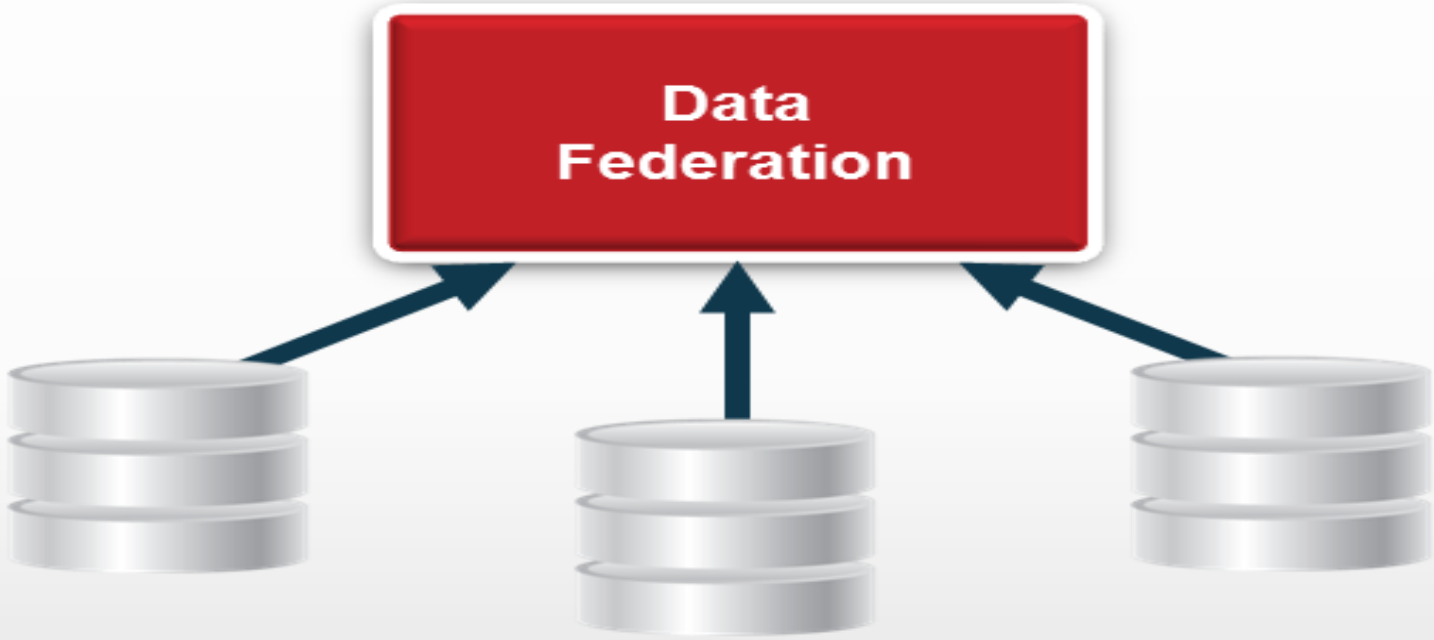
Food and Agriculture Organization of the United Nations

World Health Organization

To Support this a Data Scientist wants

- ***Self service datasets*** on demand based on metadata queries
- Virtualization of Data so that that appears to be in one place … and to query 1000s of data sources ***as a single database***.
- ***More agility than data warehousing can provide***:
  just-in-time algorithm execution,
  easy data source addition,
  live data integration and querying,
  minimize movement of bulk data.
- Run simple declarative queries, ***not program*** workflows scripts

DATA FEDERATION – *"JUST IN TIME"*



- Leave the data where it is …
- Fetch or transform it in real time … when you need it
- Integrate diverse data types it on a per query basis …
- …. with the option to publish as RDF Linked Data !

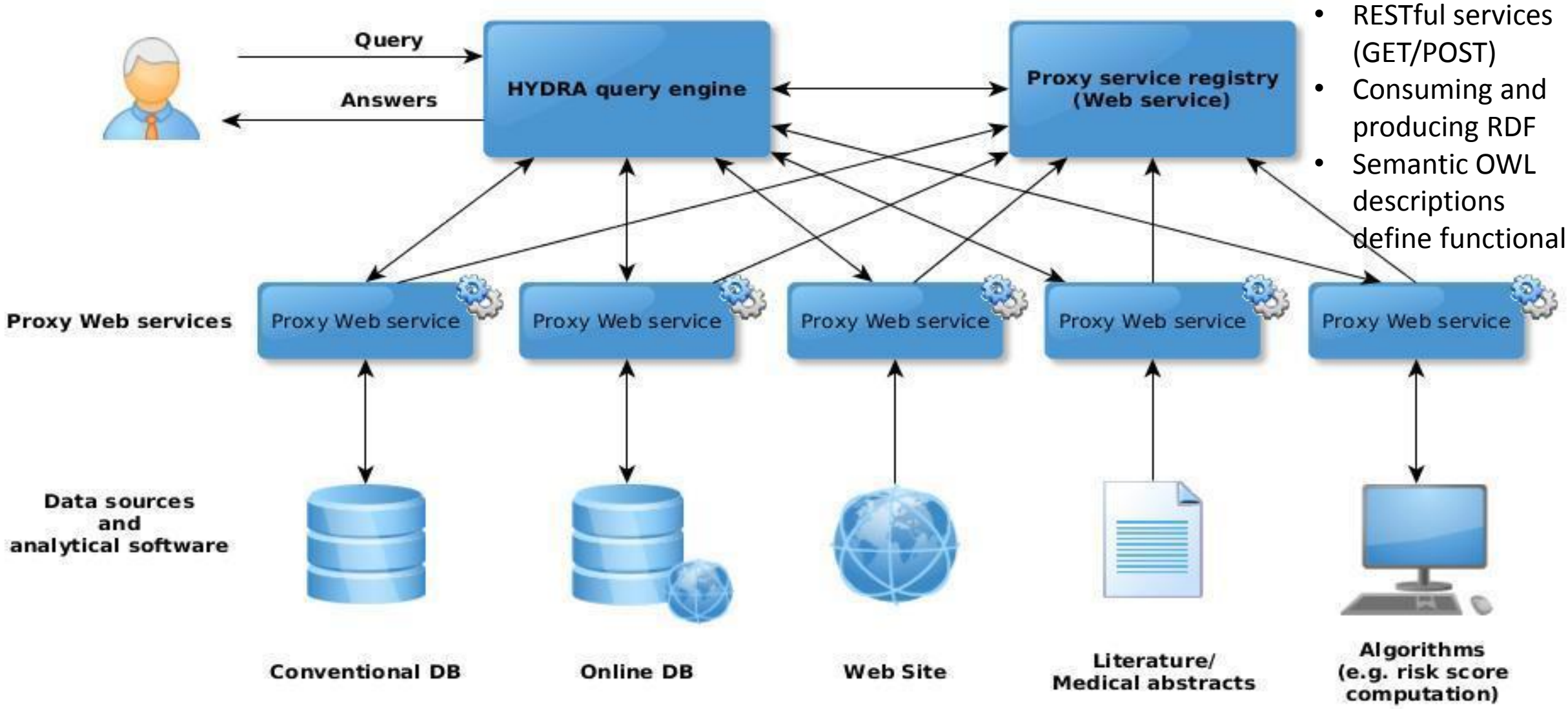
Open Semantic Web Service Middleware

Semantic Automated Discovery and Integration – *SADI

*Wilkinson etal 2009
… Just before FAIR

- A design pattern for formal description of Web Service I/O
- Service Ontology consist of I/O class names, predicates and service name and with a unique URL.
- SADI: semantic predictate describing functionality of a service is ***discoverable*** *with SPARQL* ***(GetCropDiseasebyGeoLocation)***
- RDF output of one SADI service can be consumed by another permits complex workflow composition **Syntactic Interoperability**

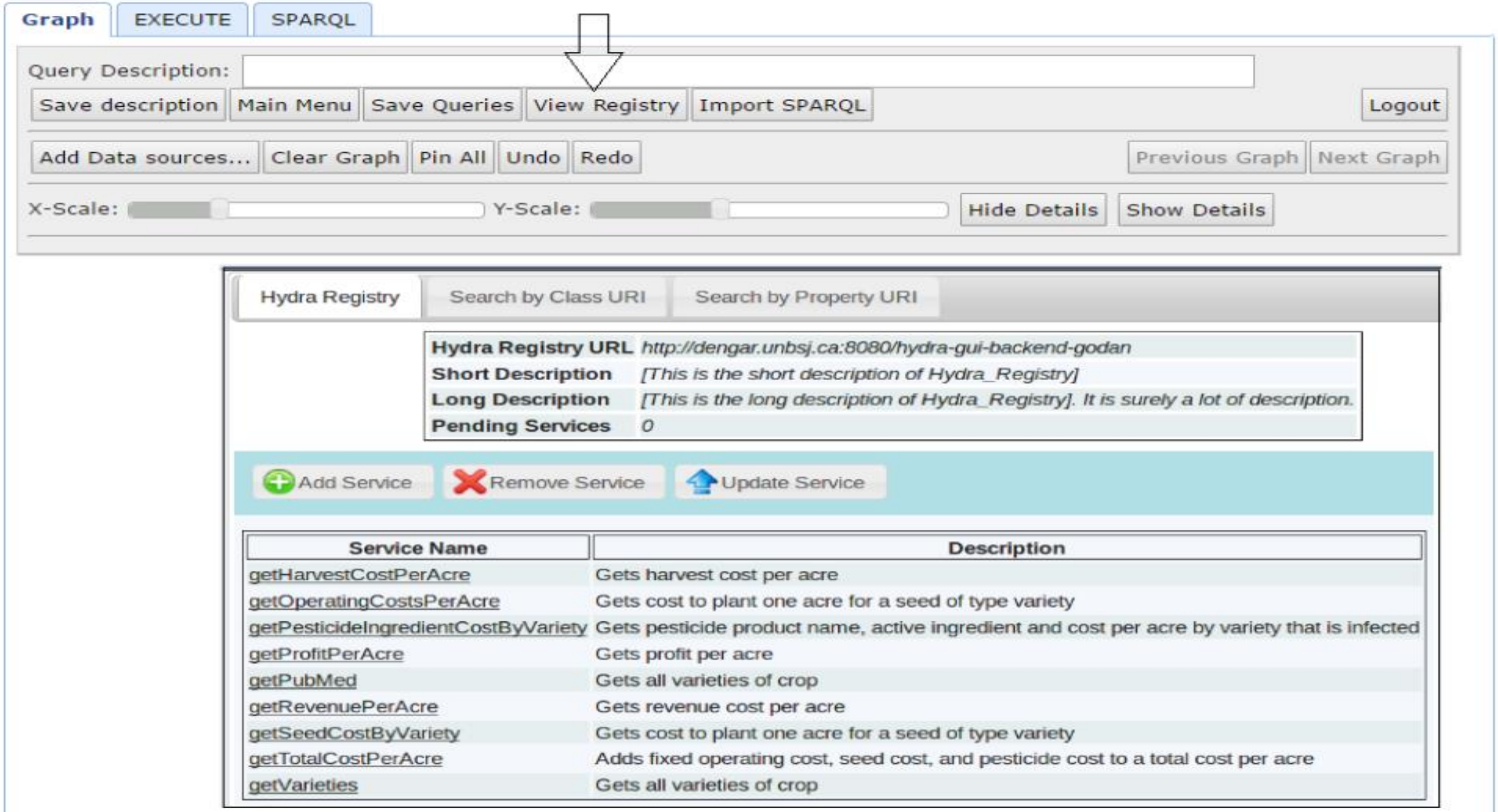
DATA FEDERATION WITH SADI SEMANTIC WEB SERVICES



- RESTful services (GET/POST)
- Consuming and producing RDF
- Semantic OWL descriptions define functionality.

100 % Syntactic Interoperability

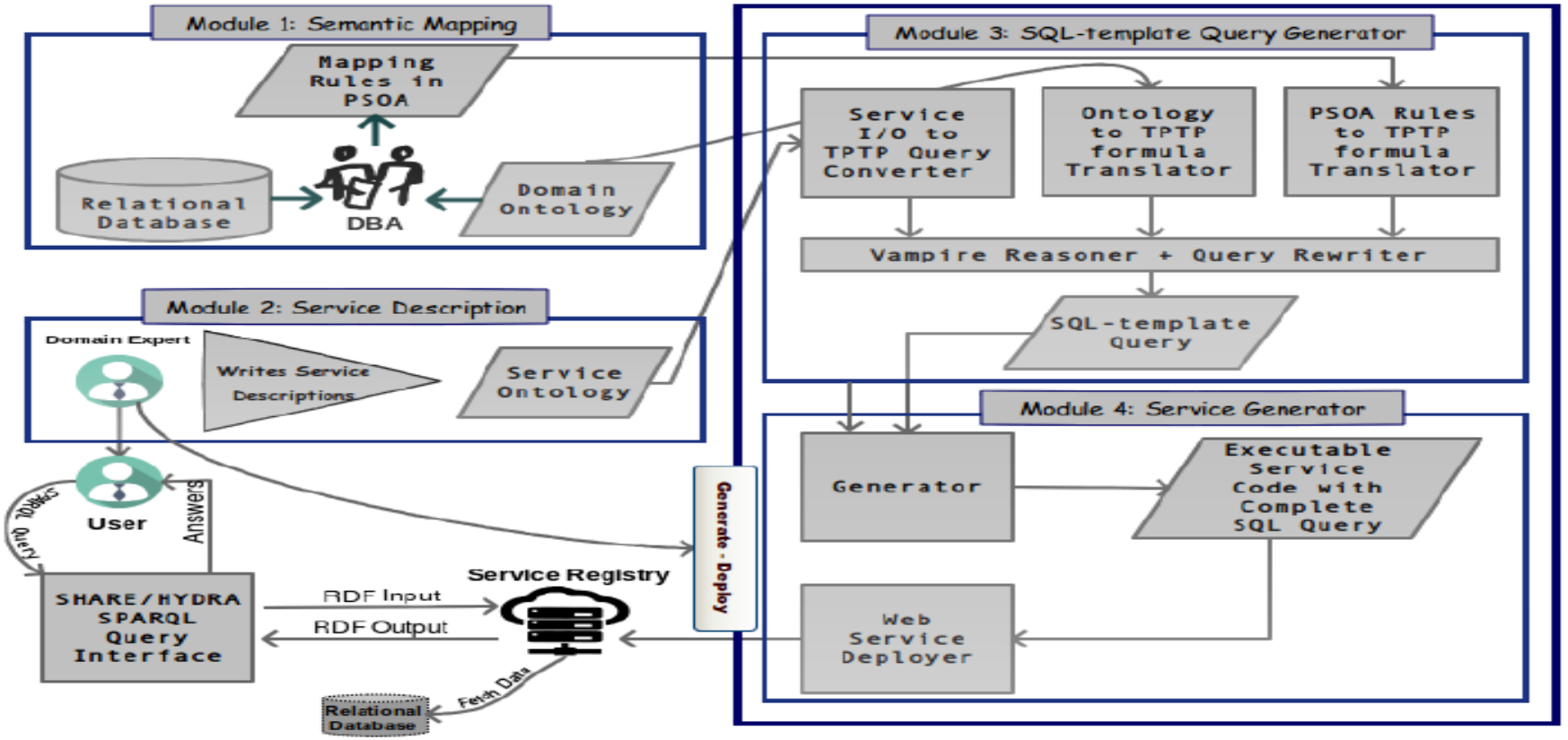
FINDABLE: MACHINE-READABLE SERVICE REGISTRY – LOOKUP !



VALET SADI

Auto generation of SADI semantic web Services (2016)

Easy addition of new services (10 seconds)



SADI Service Name	Ontology Terminologies	Type of Service	Ontology to Database Mapping Rules	ValetSADI Compatible
getHarvestCostPerAcre	User defined	Computational	Not required	No
getOperatingCostsPerAcre	User defined	Computational	Not required	No
getPesticideIngredientCostByVariety	AGROVOC	Query	Required	Yes
getProfitPerAcre	User defined	Computational	Not required	No
getPubMed	AGROVOC	Query	Required	Yes
getRevenuePerAcre	User defined	Computational	Not required	No
getSeedCostByVariety	AGROVOC	Query	Required	Yes
getTotalCostPerAcre	User defined	Computational	Not required	No
getVarieties	AGROVOC	Query	Required	Yes

http://sadiservicegeneration-rulemlapi.rhcloud.com/valetsadi/#

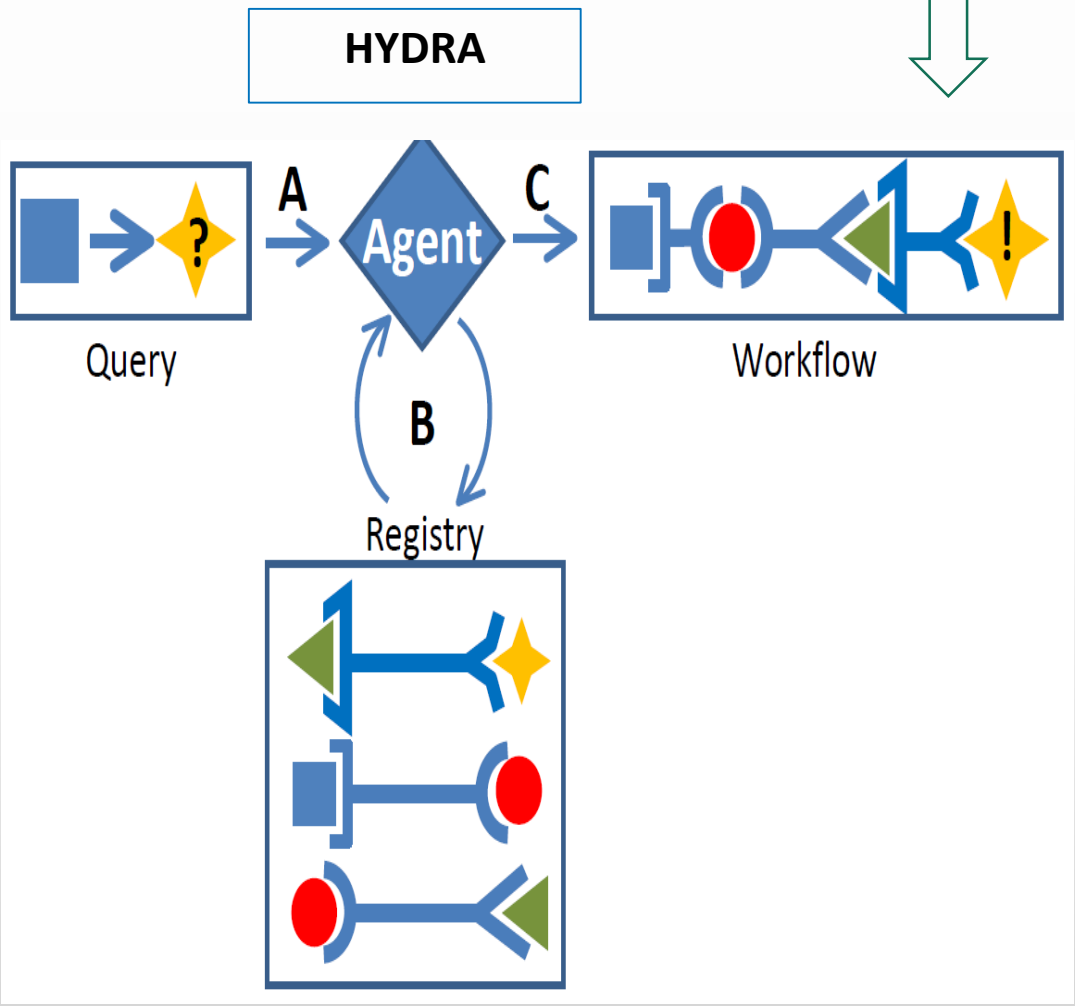
SPARQL Query Engine: HYDRA

Algorithms too !

A. **HYDRA** receives inputs in SPARQL

B. **HYDRA** can read SADI descriptions and discovers what can be computed and build a query plan.

C. **HYDRA** matches a user's SPARQL query to services, builds a workflow, orchestrates and executes queries, fetches data from disparate sources, semantically consolidates outputs.



HYDRA-BASED PRODUCTS

- HYDRA + GUI = self-service query tool
  (both standalone and cloud-based edition with a critical mass of SADI services)
- HYDRA as a Java API to be used as middleware.
- To be embedded in OEM partners' software.
- HYDRA-based turnkey solutions for
  Bioinformatics / Clinical IT / Agriculture / Finance

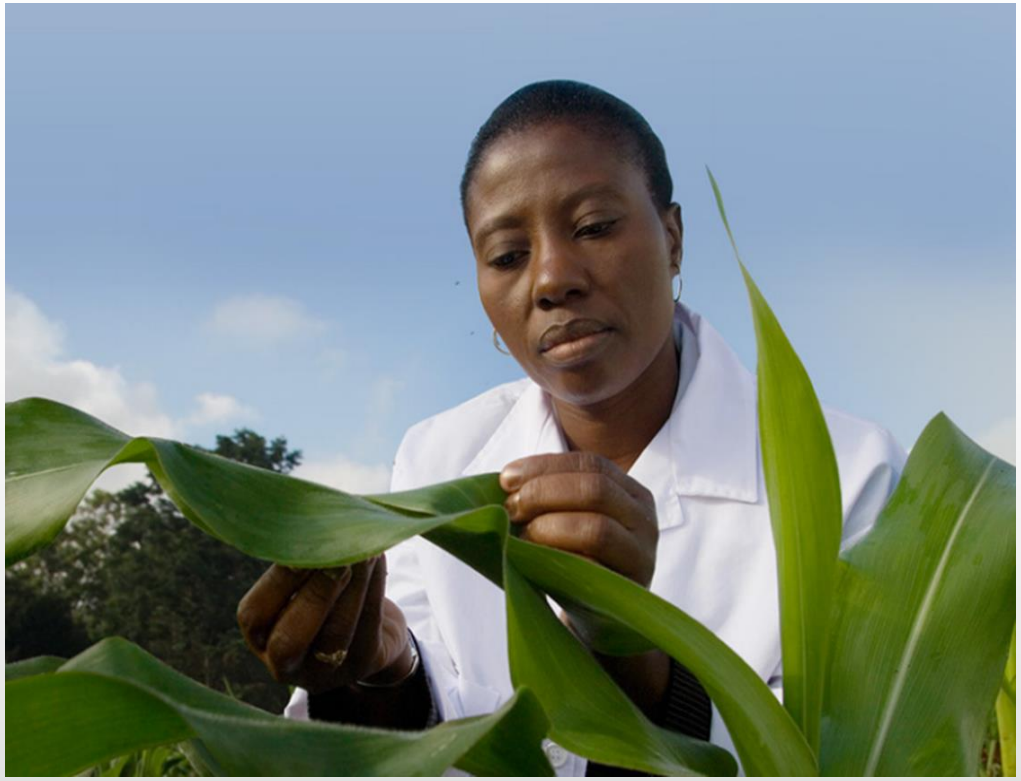
USES W3C STANDARDS: OWL / RDF / SPARQL

http://ipsnp.com/hydra/

Self-service Federated Querying

Challenges to overcome:

- Dynamically capturing users intent for agile query composition
- Feedback to the use that the query was comprehensively understood
- Conversion of a query to semantic querying language (SPARQL)

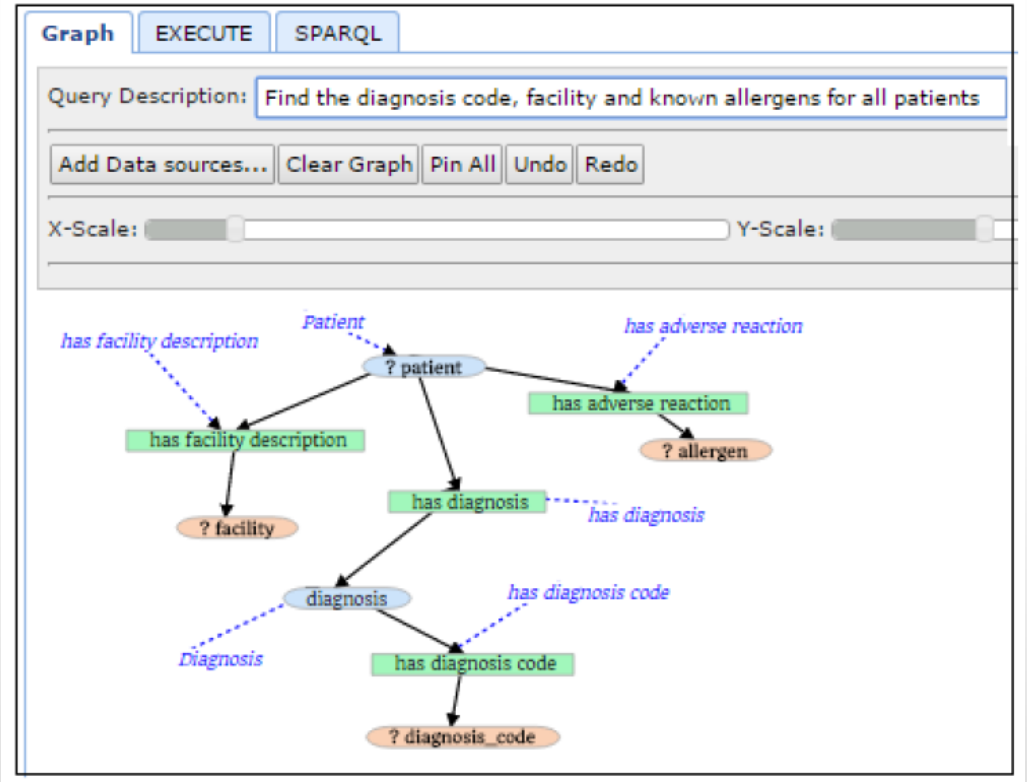


Agronomy Expert

SELF-SERVICE FEDERATED QUERYING

Challenges to overcome:

- Dynamically capturing users intent for agile query composition
- Feedback to the use that the query was comprehensively understood
- Conversion of a query to semantic querying language (SPARQL)



KEYWORD - GRAPHICAL QUERY

Story 1 What can I plant?

An agricultural consultant is helping a farmer decide what eggplant variety to plant.

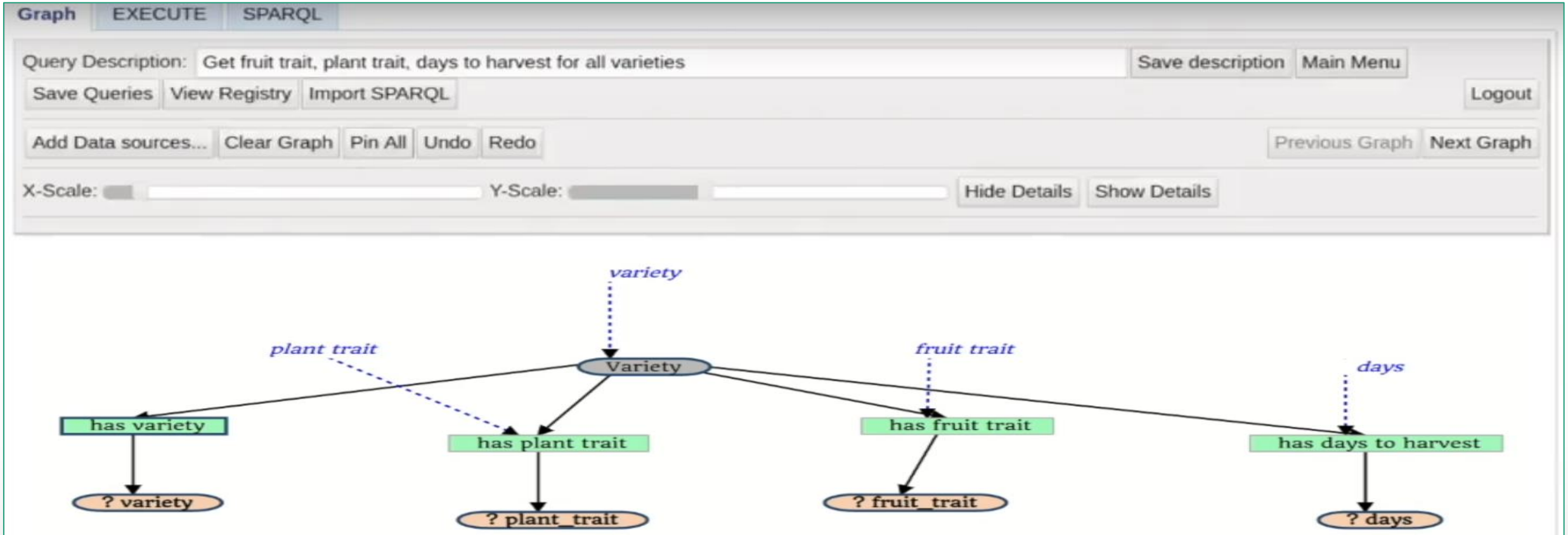
Farmer: Here are the eggplant varieties I am considering. Are there better ones?

*Consultant: enters **four** words*

variety
plant trait
fruit trait
days

Hydra: generates a tree of services using these words

Consultant: specifies that the four fields are variables in the query




```
PREFIX  ito:  <http://cbakerlab.unbsj.ca:8080/ont/ipsnp_time_ontology.owl#>
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>

SELECT  ?days ?fruit_trait ?plant_trait ?variety
WHERE
  { ?Variety  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  <http://boba.unb
              <http://boba.unbsj.ca:8080/godanipsnp/agricultural_terminology.owl#ha
              <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  <http://boba.unbs
              <http://boba.unbsj.ca:8080/godanipsnp/agricultural_terminology.owl#h
              <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  <http://boba.unb
              <http://boba.unbsj.ca:8080/godanipsnp/agricultural_terminology.owl#h
              <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  <http://boba.unb
              <http://boba.unbsj.ca:8080/godanipsnp/agricultural_terminology.owl#h
  }
```

days	fruit_trait	plant_trait	variety
67.0	Dark	Continuous Yield	Machiaw
70.0	Excellent Flavour	High Yield	Jaylo
58.0	Quick Cooking	Early Variety / Cold Tolerant / Heat Tolerant	Orient Express
60.0	Firm	Early Variety / High Yield / High Altitude	Galine
65.0	Dark fruit	High Yield	Dancer
	Light and textured	High Yield / Early Maturity	Clara

LET'S WATCH IT

https://tinyurl.com/IPSNP-HYDRA-Videos-Agric

Story 2 What is profitable?

Consultant: How much do you hope to harvest?

Farmer: Usually I can fill a 20 pound box with from one square meter of land.

Consultant: At 1.25 pesos per pound, let us look at your profits.

*Consultant: enters **four** words*

variety
price
yield
profit

Hydra: generates a tree of services using these words

Consultant: adds two fields as variables and two fields as values

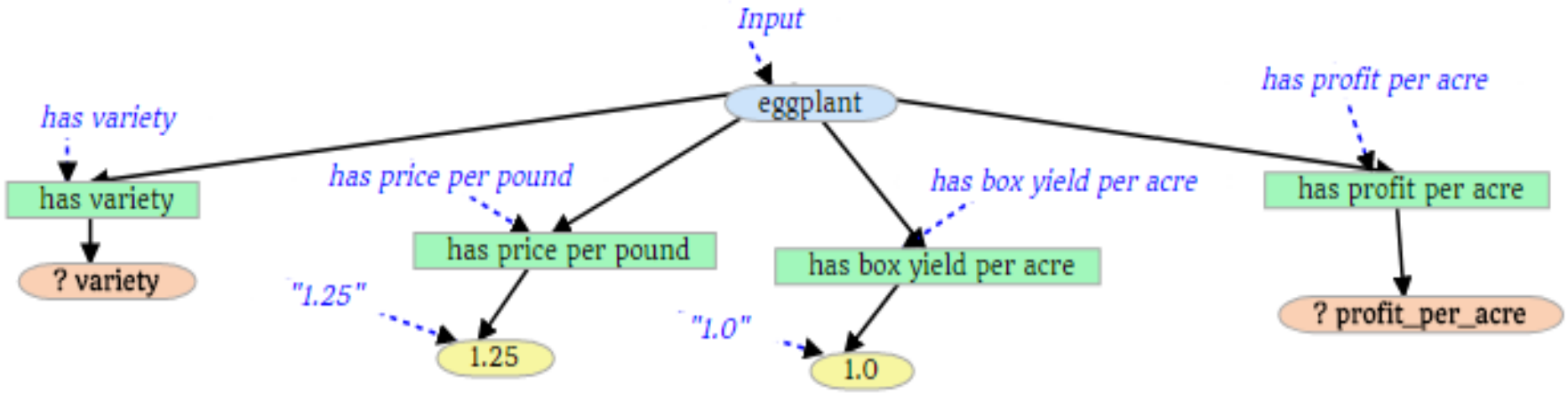
Graph | EXECUTE | SPARQL

Query Description: Get profit per acre for all varieties of eggplant | Save description

Main Menu | Save Queries | View Registry | Import SPARQL | Logout

Add Data sources... | Clear Graph | Pin All | Undo | Redo | Previous Graph | Next Graph

X-Scale: Y-Scale: | Hide Details | Show Details



profit_per_acre	variety
3887.6	Barbarella
4234.1	Dancer
4007.35	Machiaw
3607.85	Fairy Tale
3927.85	Nadia
3635.1	Clara
4026.7	Orient Express
4007.95	Galine
3602.4	TraviataOrganic
4032.0	Jaylo
4198.35	Orient Charm
3410.45	Hansel

CONSULTANT'S REPORT: SELECT VARIETY "DANCER"

- Dancer - is high yield, with dark fruit, and only 65 days to harvest.
- Dancer is a profitable variety.

profit	variety
97356.45	Orient Express
97278.9	TraviataOrganic
97208.8	Clara
96932.4	Jaylo
97692.55	Dancer

days	fruit_trait	plant_trait	variety
67.0	Dark	Continuous Yield	Machiaw
70.0	Excellent Flavour	High Yield	Jaylo
58.0	Quick Cooking	Early Variety / Cold Tolerant / Heat Tolerant	Orient Express
60.0	Firm	Early Variety / High Yield / High Altitude	Galine
65.0	Dark fruit	High Yield	Dancer
65.0	Light and textured	High Yield / Early Maturity	Clara

LET'S WATCH IT

https://tinyurl.com/IPSNP-HYDRA-Videos-Agric

Preparing the Profit Service

- Profit is the expected revenue minus expected costs.
- Revenue: expected market price of the eggplant, and the expected yield in boxes per acre.
- Costs: the total operating costs, seed and pesticide costs per acre

Production Practices and Sample Costs to Produce Eggplant / UC Davis '97

	Y(B/A)	
	1000	1100
Operating costs/Acre		
Preplant Cost	1036	1036
Cultural Cost	2184	2184
Harvest Cost	4205	4626
Postharvest Cost	100	100
Interest on operating capital	206	211
Total Operating Costs/Acre	7731	8156
Total Operating Costs/Box	7.73	7.41
Total Operating Costs/Box	7.73	7.41
Cash Overhead Costs/Acre	228	228
Total Cash Costs/Acre	7959	8384
Total Cash Costs/Box	7.96	7.62
Non-cash Overhead	142	142
Total Costs/Acre	8101	8527
Total Costs/Box	8.10	7.75

Table 1: Costs per acre at varying yields to produce eggplant, Riverside country-1997, Ranging analysis, U.C. Cooperative extension. [Y(B/A): Yield(Box/Acre)]

Price	Y(B/A)	
($ Per Box)	1000	1100
6.25	-1481	-1281
6.75	-981	-731
7.25	-481	-181
7.75	19	369
8.25	519	919
8.75	1019	1469
9.75	2019	2569

Table 2: Net returns per acre above operating costs for eggplant. [Y(B/A): Yield(Box/Acre)]

getProfitPerAcre

Computing Profit = (total.revenue.per.acre - total.cost.per.acre) * acreage

Relies on other ***computing Services***

- getRevenuePerAcre

  total.revenue.per.acre = price.per.box * yield.boxes.per.sq.meter * sq.meter.per.acre

- *getTotalCostPerAcre (aka getSeedCostPerAcre +getOperatingCostPerAcre +getPesticideIngredientCostPerAcre)*

  total.cost.per.acre = pest.cost.per.acre + seed.cost.per.acre + preplant.cost.per.acre + cultural.cost.per.acre +postharvest.cost.per.acre + harvest.cost.per.acre + cash.overhead.cost.per.acre + noncash.overhead.cost.per.acre

getSeedCostByVariety

Biz Logic: ***Data retrieval*** from tables in the eggplant database and uses the following Java code that creates an SQL statement:

```
String queryText = "SELECT ‘Cost.per.acre‘"
+ " FROM eggplantseed"
+ " WHERE ‘Eggplant.Variety‘ = \"" + INPUT + "\"";
```

The database table eggplant seed has two columns: Eggplant.Variety and Cost.per.acre. The service accepts a variety name as INPUT, passes it to the SQL engine, and returns the cost of seed for one acre planted.

Pesticide impact on bees (data from Literature)

https://tinyurl.com/IPSNP-HYDRA-Videos-Agric

Story 3 Pesticides and Bees

Farmer: If I treat the pests, will I harm my bees?

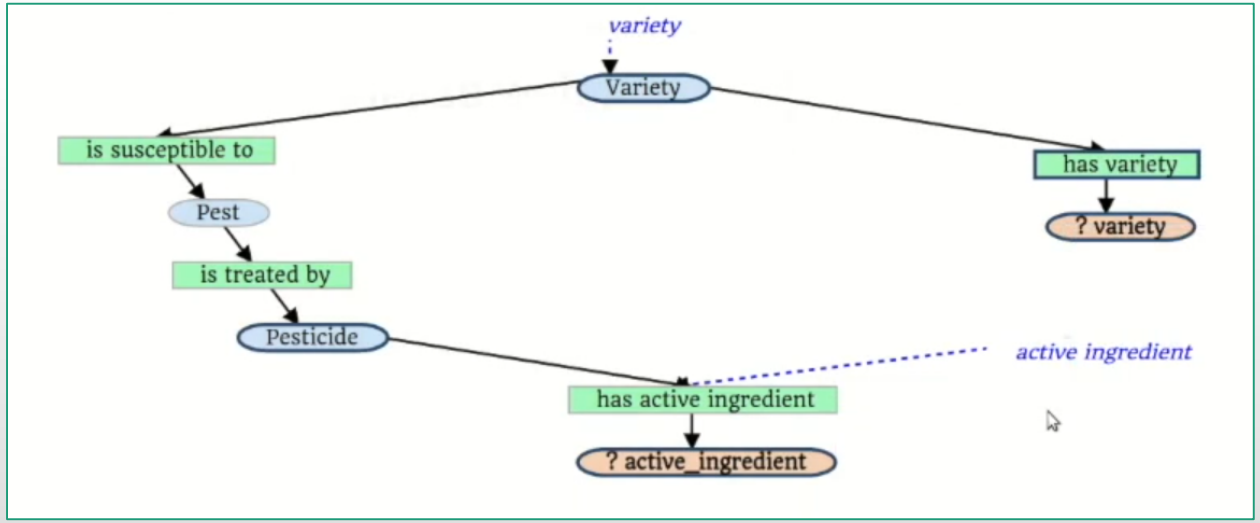
Consultant: Let us look at the common pets for each variety and the pesticides for each one

*Consultant: enters **two** words*

variety
active ingredient

Consultant:

specifies two fields are variables



Story 3 Pesticides and Bees (continued)

Consultant: OK. Now let us look at the effects on non-target organisms including bees

*Consultant: enters **five** words*

pubmed

active ingredient

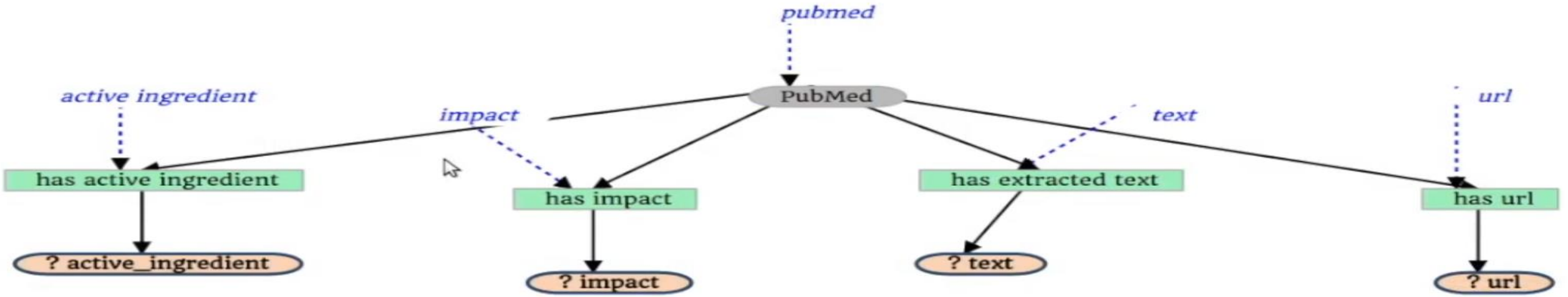
impact

text

url

Consultant: specifies four fields are variables

HYDRA GENERATES A QUERY TREE




```
PREFIX  ito:  <http://cbakerlab.unbsj.ca:8080/ont/ipsnp_time_ontology.owl#>
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>

SELECT  ?active_ingredient ?impact ?text ?url
WHERE
  { <http://localhost:8080/godanipsnp/agricultural_terminology.owl#PubMed>
          <http://localhost:8080/godanipsnp/agricultural_terminology.owl#has_active_ingredient>  ?active_ingredient ;
          <http://localhost:8080/godanipsnp/agricultural_terminology.owl#has_extracted_text>  ?text ;
          <http://localhost:8080/godanipsnp/agricultural_terminology.owl#has_url>  ?url ;
          <http://localhost:8080/godanipsnp/agricultural_terminology.owl#has_impact>  ?impact .
  }
```

CONSULTANT'S REPORT: ACETAMPRID IS SAFE

Lethal effect concentrations of neonicotinoids for workers of the honey bee (*A. mellifera*) by oral and contact exposure as determined in different laboratory studies

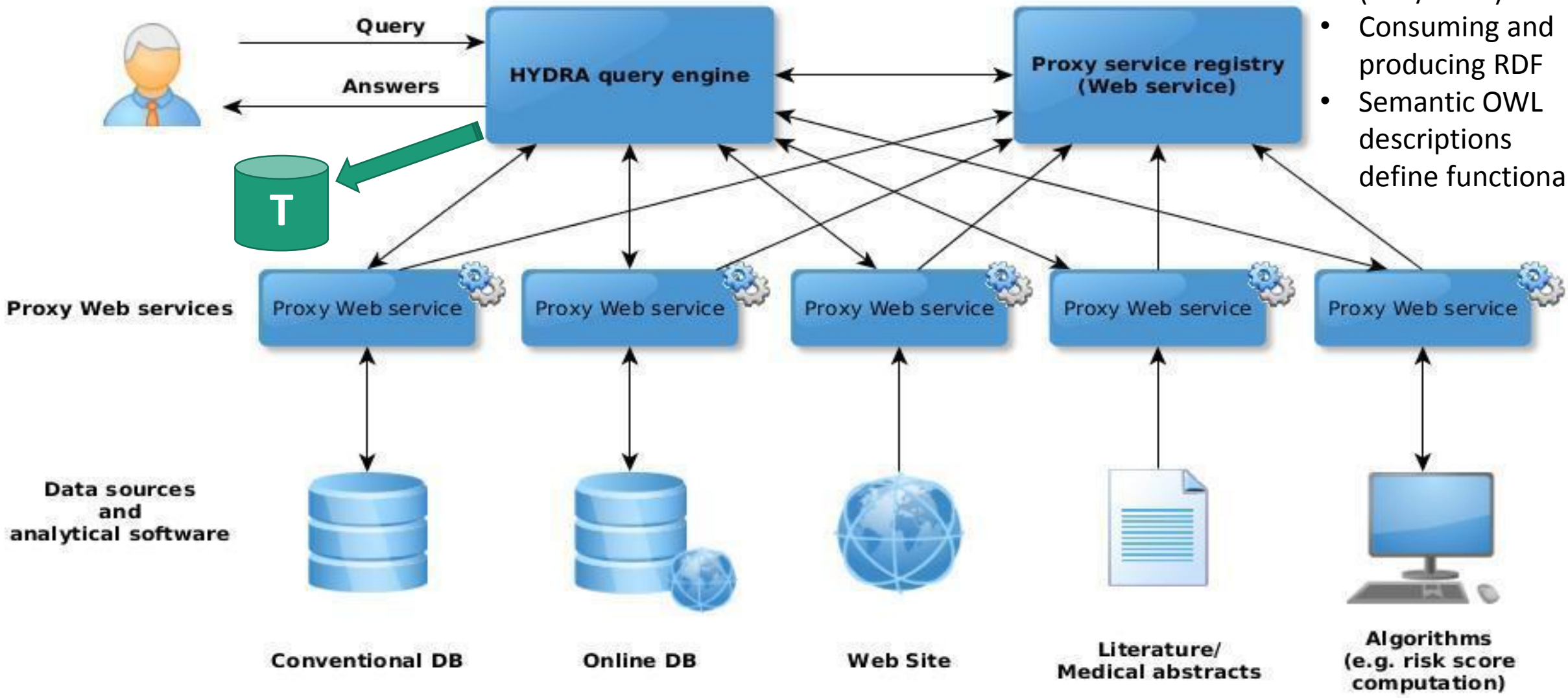
Neonicotinoid	Exposure	LD_{50} (µg bee^{-1})
Parent compound		
Acetamiprid	Contact: individual bee (acute; no info on concentration range)	24 h: 7.07
Acetamiprid	Contact + oral: individual bee (acute; no information on concentration range)	48 h: 14.5 (oral) + 8.09 (contact)
Acetamiprid	Contact to dry residue + oral: 100 mg l^{-1} (acute; 2 days exposure for contact and 3 days for oral exposure)	Harmless

- Acetamiprid is harmless to honey bees
- Acetamiprid is the active ingredient in pesticides for these varieties:
  - ✓ Clara
  - ✓ Orient Express
  - ✓ Jaylo
  - ✓ Dancer

active_ingredient	variety
abamectin	Galine
acetamiprid	Clara
abamectin	Orient Charm
acetamiprid	Orient Express
abamectin	Angela
acetamiprid	Jaylo
acetamiprid	Dancer

DATA FEDERATION WITH SADI SEMANTIC WEB SERVICES

- RESTful services (GET/POST)
- Consuming and producing RDF
- Semantic OWL descriptions define functionality.



REFLECTIONS

Open Question …. Does it scale ?

Why is my SPARQL query so slow when I use prope... | Oracle Community
https://community.oracle.com/thread/3639765
Dec 11, 2014 - 3 posts - 2 authors
This is my query: PREFIX rdfs: PREFIX rdl: PREFIX xsd:

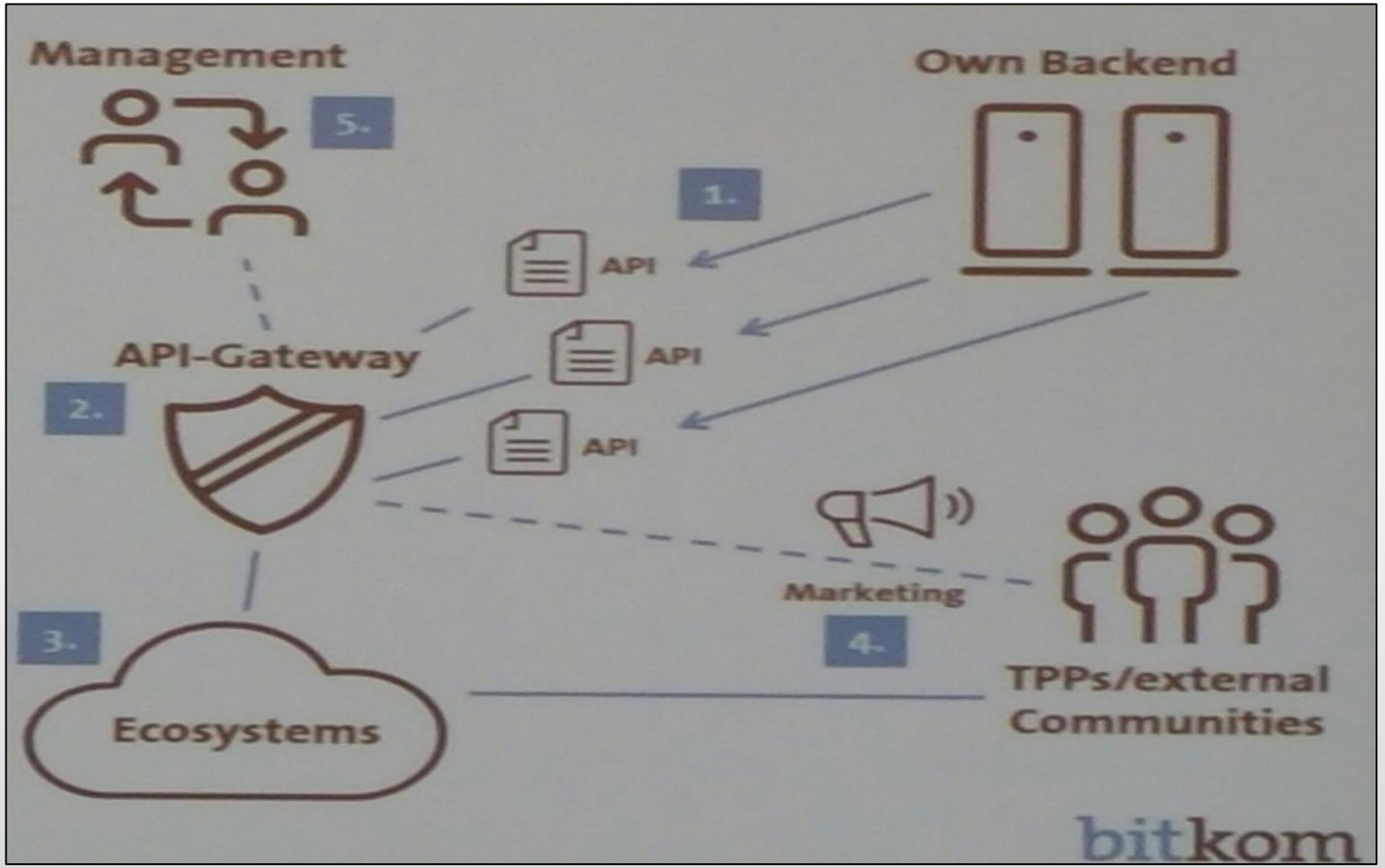
rdf - Why is my SPARQL query so slow? - Stack Overflow
https://stackoverflow.com/questions/33540751/why-is-my-sparql-query-so-slow
Nov 5, 2015 - There doesn't look to be anything particularly wrong with your **SPARQL** query and you have made no obvious mistakes (other than some ...

marklogic - Is there any way to optimize SPARQL queries? - Stack ...
https://stackoverflow.com/questions/.../is-there-any-way-to-optimize-sparql-queries
Jun 20, 2016 - I think the biggest problem is going to be the BIND() - MarkLogic 8 doesn't optimize the pattern you're using at all well. Can you try substituting ...

performance - How to make my sparql query with regex faster? - Stack ...
https://stackoverflow.com/questions/.../how-to-make-my-sparql-query-with-regex-fast...
Nov 26, 2012 - Any query using REGEX will almost certainly be **slow** unless your query restricts to a small enough portion of the dataset. Processing a REGEX ...

An API Gateway: Data for developers

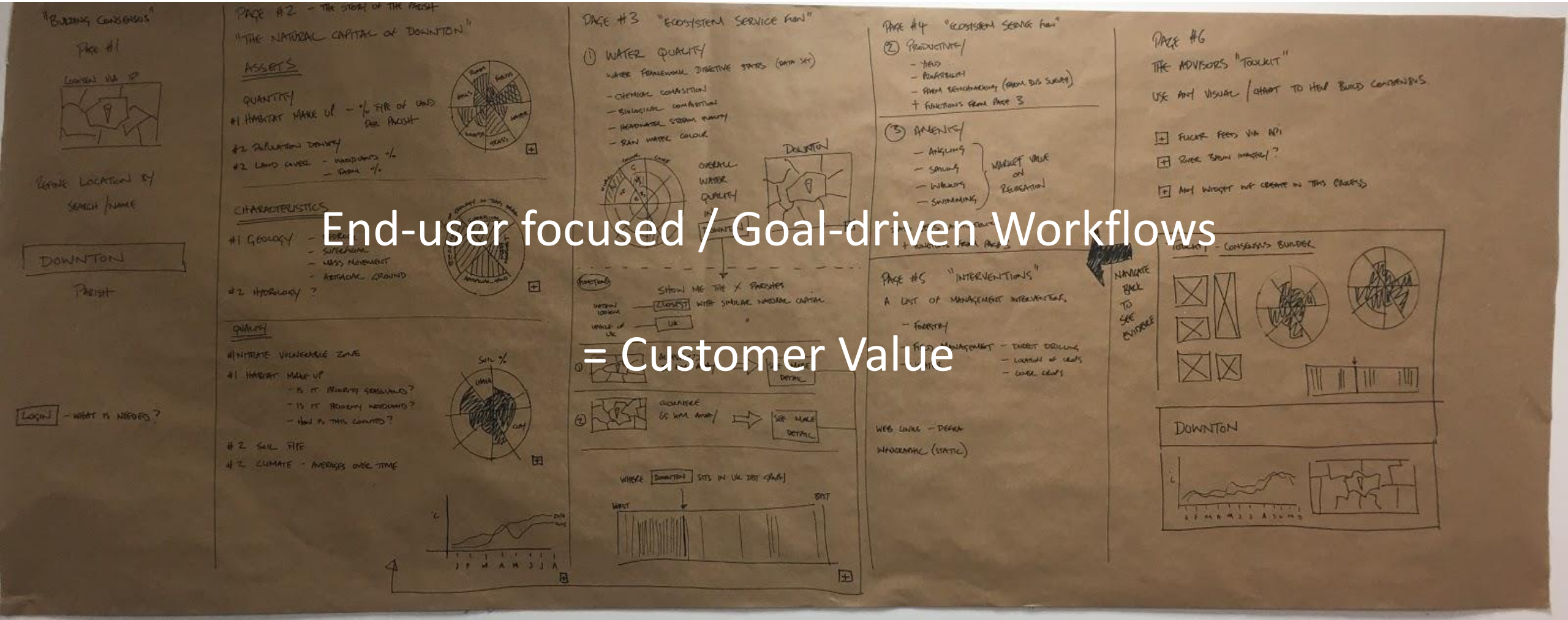
... access for developers (Only)?



Defra

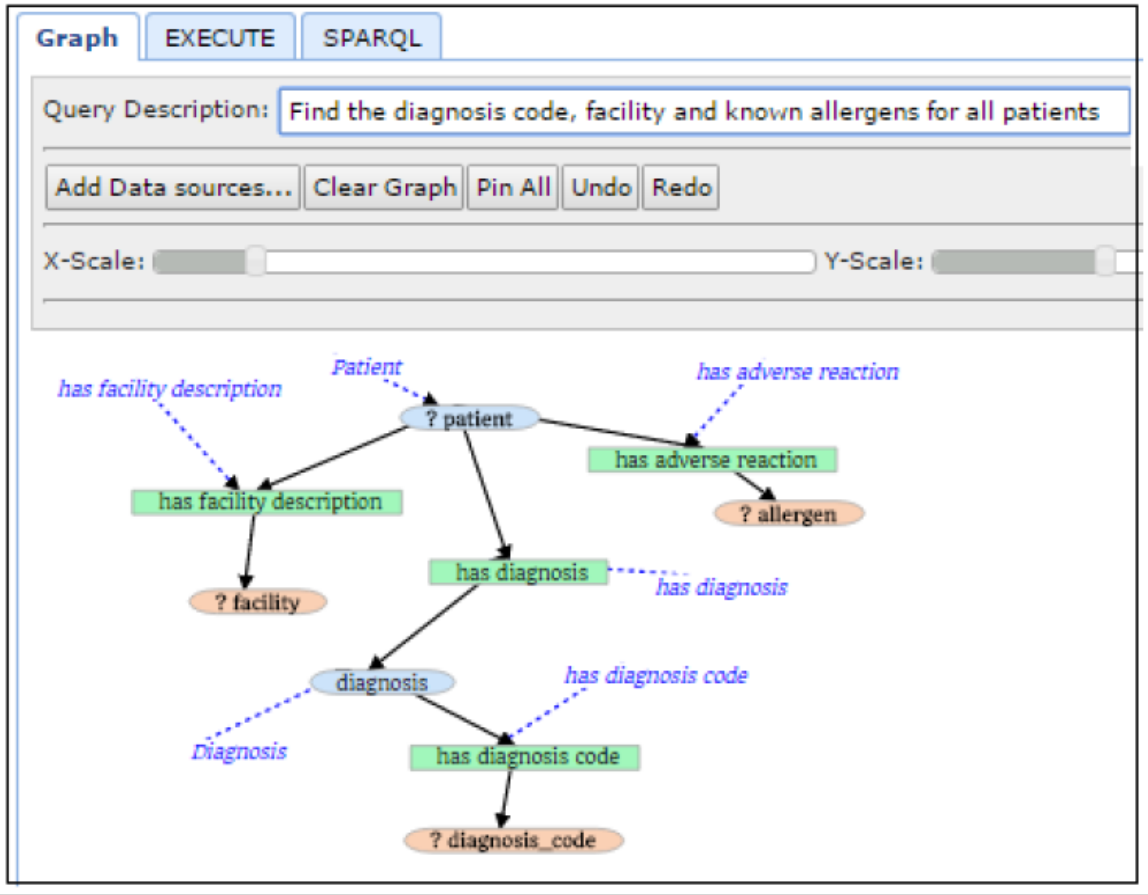
Natural Capital Explorer

agri metrics

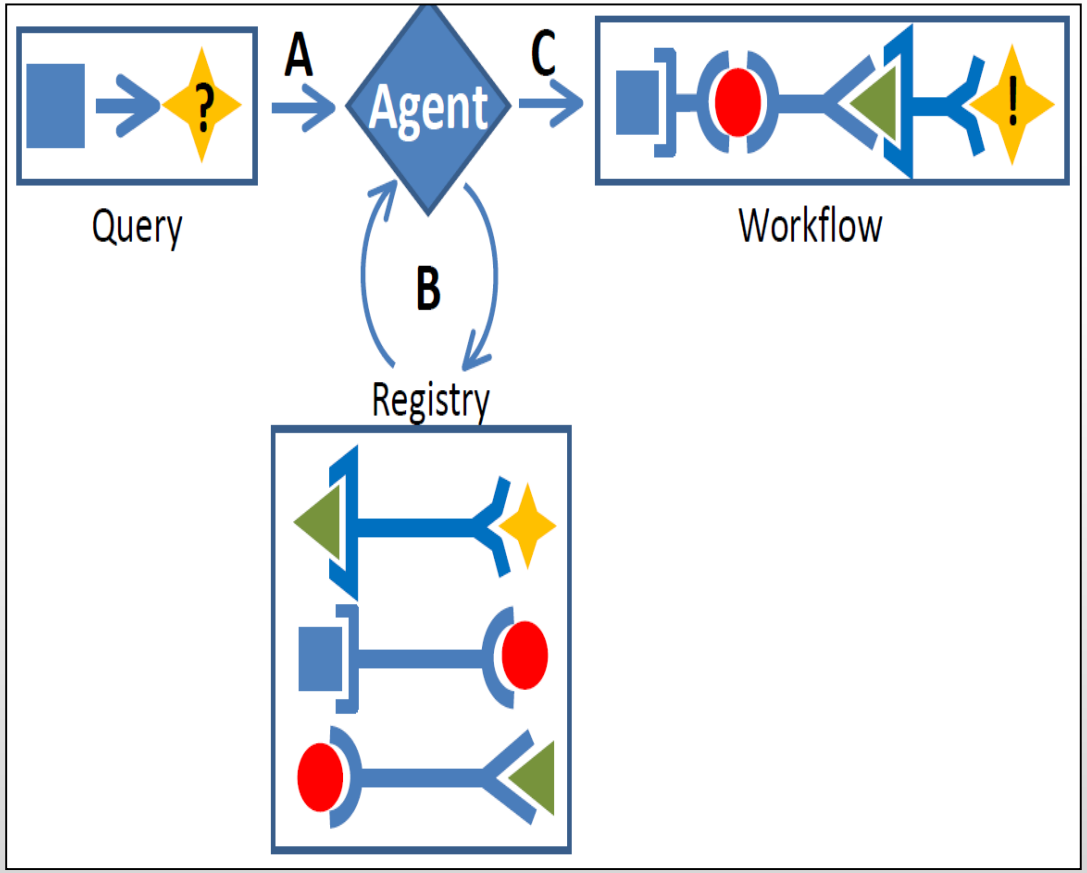


SELF-SERVICE SEMANTIC QUERYING

KEYWORD – GRAPHICAL QUERY



AND HYDRA !



CHANGE MANAGEMENT

CHANGE MANAGEMENT FOR MALARIA SURVIELLANCE

Global
Grand Challenges

BILL & MELINDA GATES foundation

ABOUT | CHALLENGES | AWARDED GRANTS | NEWS | GRANT OPPORTUNITIES

Home / Awarded Grants

A Semantic Framework to Support Evolution and Interoperability

GRAND CHALLENGES EXPLORATIONS | MALARIA ANALYTICS | 1 NOV 2016

Arash Shaban-Nejad of the University of Tennessee Health Science Center in the U.S. will develop an analytic framework to help integrate dynamic surveillance data from multiple sources and health systems to support decision making for malaria elimination. Data on malaria is currently scattered in different formats across diverse organizations, making it difficult to access and use. An ontology is a web-based method that explicitly defines specific concepts using logical rules and constraints, and can be used to capture and combine information from numerous sources into a formal framework. They will analyze existing malaria data sources such as Mapping Malaria Risk in Africa (MARA) across selected African countries with different languages including Uganda and Gabon, and use it to generate a service ontology that enables data integration, and implement a semantic web service that can also be used to access the data.

A Change Management Dashboard for the SIEMA Malaria Surveillance Infrastructure

Jon Haël Brenas[1], Mohammad Sadnan Al-Manir[2], Christopher J. O. Baker[2], and Arash Shaban-Nejad[1]

[1] The University of Tennessee Health Science Center- Oak Ridge National Laboratory Center for Biomedical Informatics, Department of Pediatrics, Memphis, Tennessee, USA
{jhael, ashabann}@uthsc.edu
[2] Department of Computer Science, University of New Brunswick, Saint John, Canada
{sadnan.almanir, bakerc}@unb.ca

Abstract. Malaria is an infectious disease that remains a major cause of death in low-income developing countries. The World Health Organization (WHO) has set a target for its eradication by 2030. Among the issues that will have to be solved to achieve this goal is interoperability between the various malaria data sources. This can be achieved through the adoption of semantic web service infrastructure to provide access to the data while abstracting its structure. Given that data sources, semantic metadata descriptions and ontologies evolve over time, it remains a challenge to propagate changes, ensuring services continue to be discoverable, while at the same time keep the services operational. We propose a dashboard to detect, identify, and classify changes based on their likely functional impact on data access, and propose steps to maintain infrastructure, either rebuilding or retiring services from a registry.

Search...

- Dashboard
- Domain Ontology
- Service Ontology
- Relational Database
- Impact on Services

SIEMA - Change Manager's Dashboard

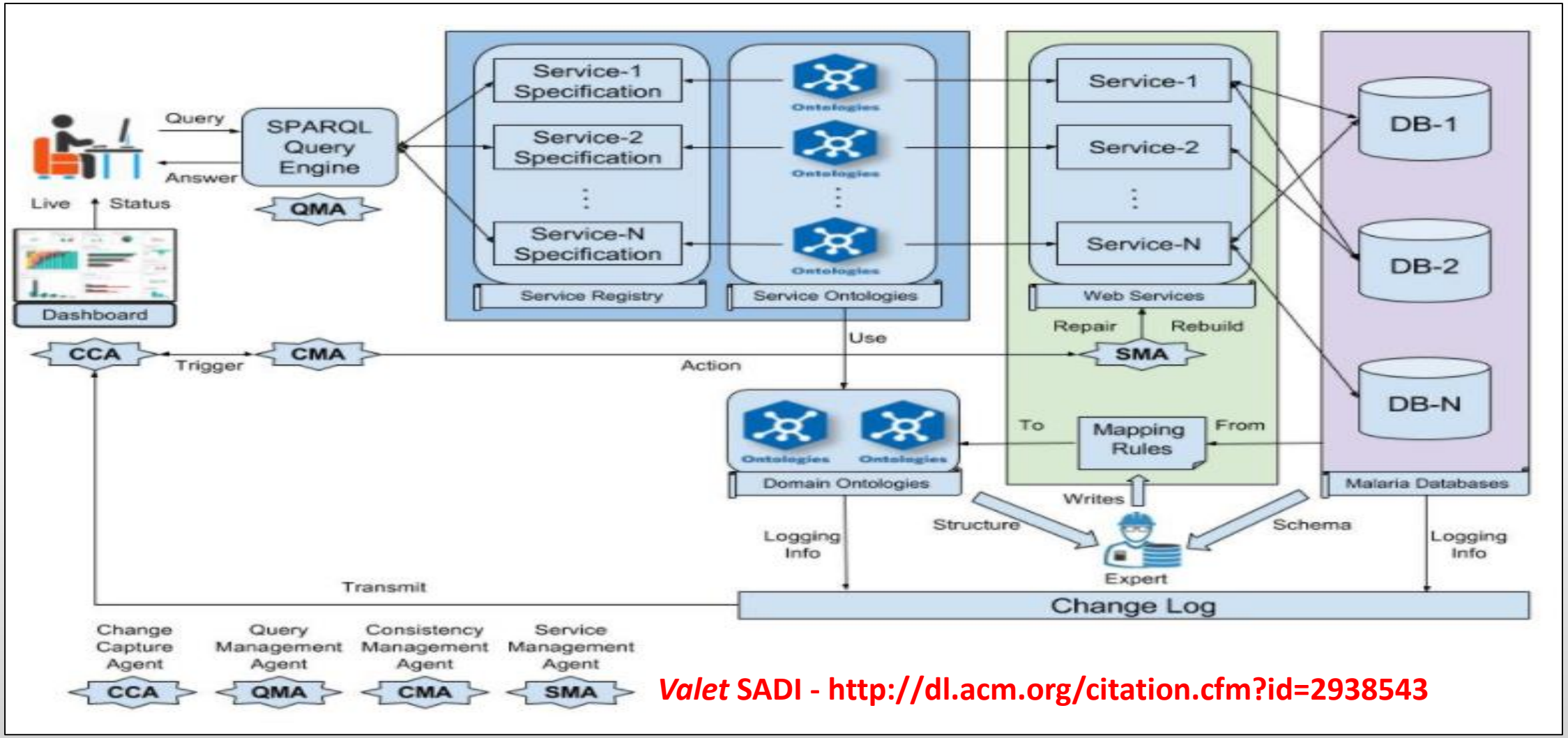
2	2	4	3
Domain Ontology	Service Ontology	Relational Database	Impact on Services
View Details	View Details	View Details	View Details

Notifications Panel

A service became inactive	*4 minutes ago*
An attribute name changed	*12 minutes ago*
An object property deleted	*22 minutes ago*
A class added	*32 minutes ago*
An object property deleted	*41 minutes ago*

View All Alerts

SIEMA: Malaria Analytics Quality Control and Change Management with SADI Services



Valet SADI - http://dl.acm.org/citation.cfm?id=2938543

THANK YOU

Acknowledgements

Dr Robert W. Caudwell

Applied Research and Innovation Systems in Agriculture Project (ARISA) funded by the Commonwealth Scientific and Industrial Research Organization of Australia (CSIRO)

Traction in the World of Big Data

iPSNP