



[Merck IT: Innovation]

Exploring Semantic Web
Technologies within Research
and Development



Jaime Melendez
Technology Innovation
Merck and Co., Inc.

Presented at C-SHALS 2009
26-FEB-2009

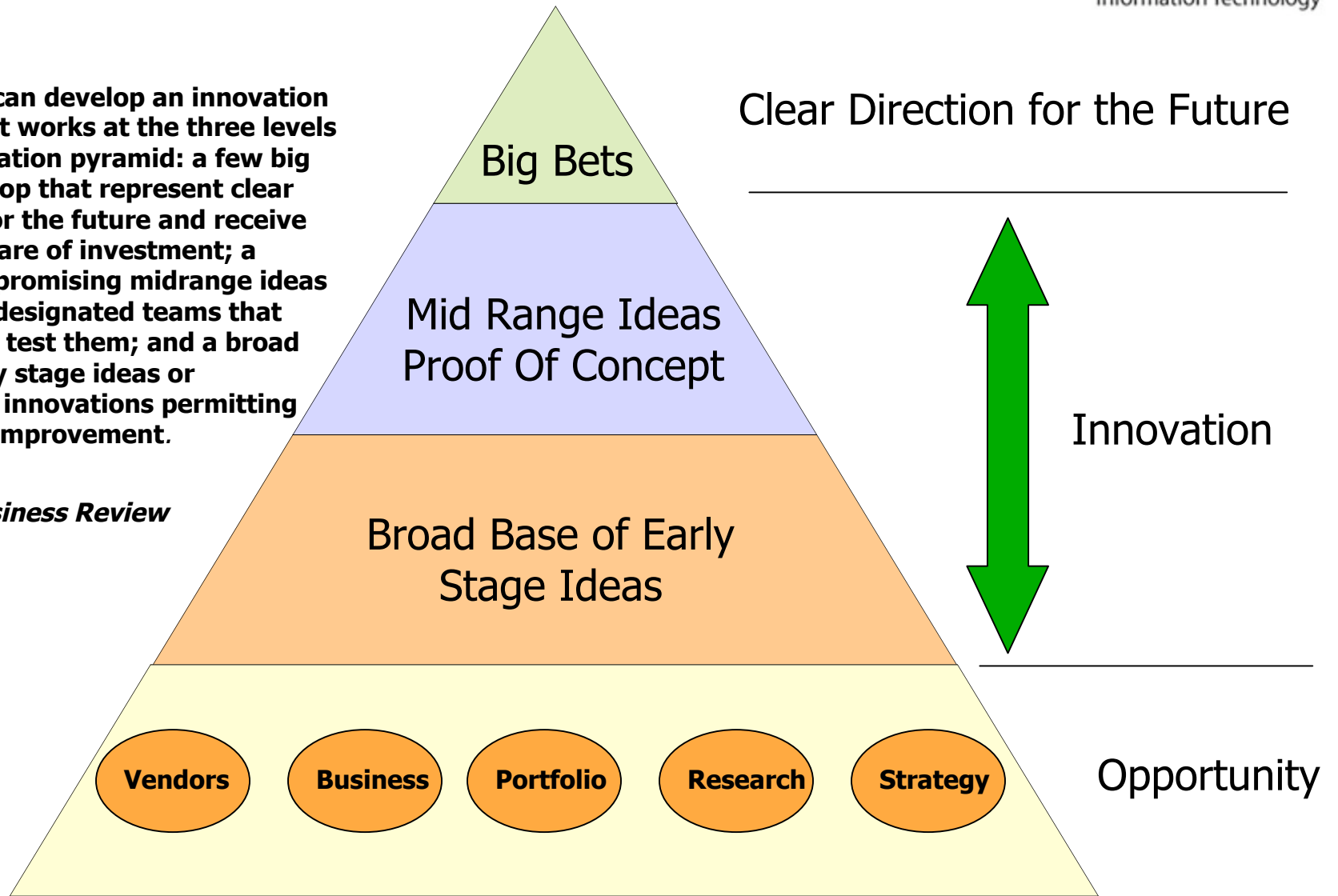
Agenda

- Overview of Innovation Process
- Semantic Web Technologies Proof-of-Concept
 - Background / Description
 - Challenges
 - Experiment Analysis
 - Conclusions
- Q&A

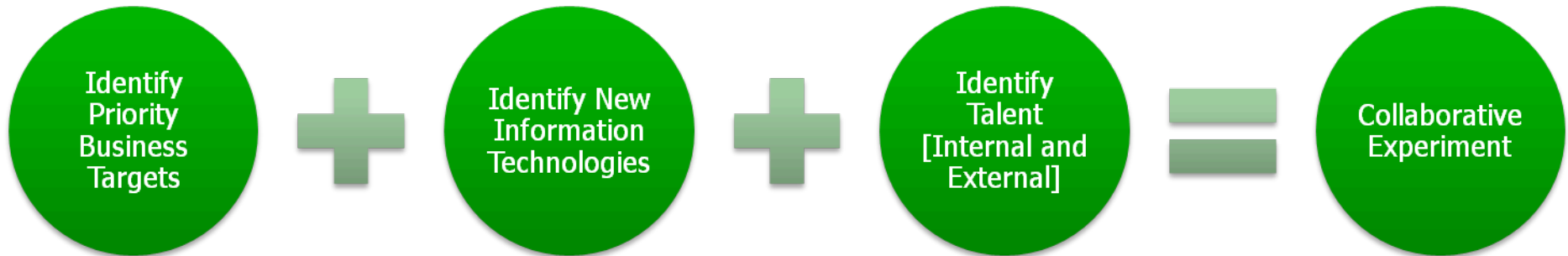
Why Innovate?

Companies can develop an innovation strategy that works at the three levels of the innovation pyramid: a few big bets at the top that represent clear directions for the future and receive the lion's share of investment; a portfolio of promising midrange ideas pursued by designated teams that develop and test them; and a broad base of early stage ideas or incremental innovations permitting continuous improvement.

*R.M Kanter
Harvard Business Review*



Overview of Innovation Process



- Demonstrate capabilities, value and feasibility
 - Provide Go / No Go decision

Life in the R&D Lab

- Scientists need a system for storing and accessing lots of biologics data records
- The data structure is changing rapidly due to additions of new biologics properties, and new relationships between molecular entities.
 - Prevents easy RDBMS implementation
- Scientists also need an easy way to configure views for navigating and looking at different sets of data
- Scientists also want to infer new insights from data, rather than simply query it.
- They also desire the ability to make annotations in such a way that they can be also used during inferencing.

Challenges of Ever-Changing R&D

- Integration of diverse sources of data
- Unanticipated data requirements after implementation
- Improving quality control on large datasets
- Adhering to enterprise standards



Semantic Web: Identified Benefits

- Shared, yet controllable, data access
 - Implements established and emerging standards
- Improved knowledge management
 - Automating the process of capturing knowledge
 - Machine data mining to produce insights not seen before
- Flexible data model
 - Ontology evolves around the data
- Quick prototyping and development
 - Potential low cost via open-source
 - RDF output is independent of the vendor/application

Semantic Web Technologies

Proof-of-Concept

SW Project Overview

- Developed functional prototype, using standard W3C Semantic Web technologies, to allow Biological area to integrate antibody data into semantic web form
- PoC provides an integrated and searchable data repository supporting Biological area
- Used Agile Development method to develop PoC in under 3 months

Benefits

- Provides flexible data integration accommodating diverse data sources
- Reduces manual, intensive, and error-prone data processing resulting in reduced data integration costs and risks

Outcome Overview

- PoC Semantic approach provides capabilities not provided with traditional integration strategies:
 - Ability to formally annotate findings and interpretations
 - Ability to search on term meaning and prioritize those results
 - Ability to infer new insights
 - Ability to improve library design as a result of a flexible data model

Developing the PoC

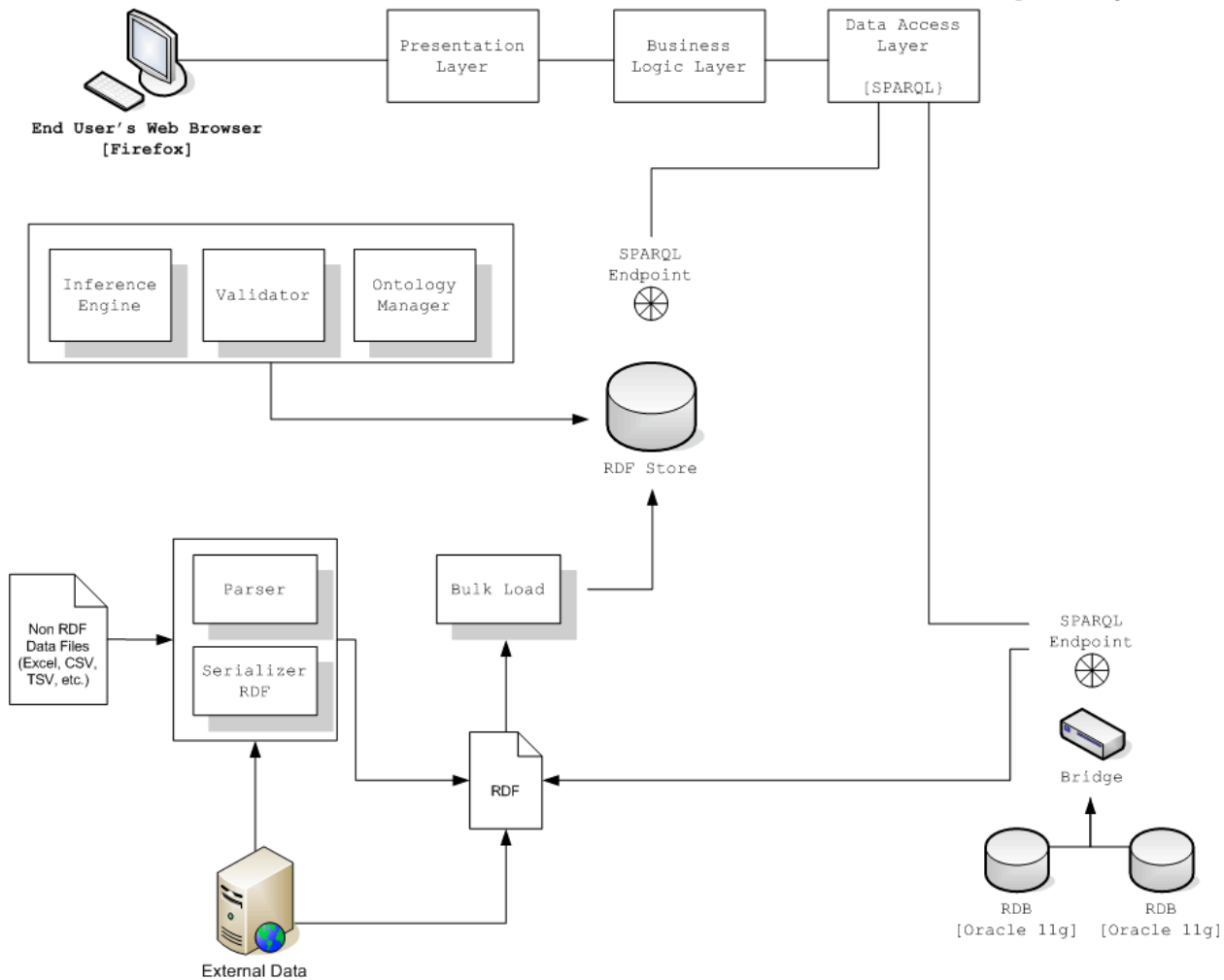
- Agile development method
 - 10 week PoC project
- Approximately 8 weeks of application development time
 - Iteration 1 – duration 3 weeks
 - User Interface/Concept-Based Searching
 - Iteration 2 – duration 2 weeks
 - Content Annotation and Query Builder
 - Iteration 3 – duration 3 weeks
 - Inference and Rule Builder
- Application Development Team
 - UI Developer
 - Subject Matter Expert
 - Project Manager/Java Developer

Designing the Information Model

- Ontology developed using open source Protégé
 - 2 weeks
- Information Model Development Team
 - Semantic Web Ontology Model Expert
 - Business Area Representative
 - Project Manager
- Modifications to ontology model are easily implemented
 - Flexibility supports dynamic business area

System Architecture

[STB System View]



Iteration 1: User Interface/Concept-Based Searching

■ Intent

- Provide a precise and concept-aware search capability to span across multiple knowledge sources, both structured and unstructured.

■ Business Benefits/Results

- Improved search precision
- Improved expressiveness

■ Technical Outcome

- Reusable data parser to convert excel data into RDF
- Expandable User Interface

Semantic Web Technology for
MRL Biologic POC [STB]

\$Date: 09-JUL-2008 10:16:28#\$ \$Revision: 219 \$

Welcome Data Loader Faceted Oligo Viewer Faceted Sequen

Data Loader

File Type Oligo

Oligo
Sequence

Browse... Upload

Uploading

Iteration 2: Content Annotation and Query Builder

- **Intent**
 - Provide ability to add annotations to electronic content in order to capture decision rationale, work comments and discussions
- **Business Benefits/Results**
 - Improved capture of decision rationale, better ability to collaborate remotely
- **Technical Outcome**
 - Annotation implementation can be applied to any (URI-specified) data, even outside the current project
 - Query Builder empowers users to develop custom data retrieval queries

Annotation saved successfully

Submit Annotation

1 oligos found, displaying 1 oligos, from 1 to 1. Page 1 of 1.

name	has_seqnum	length	MW
AL-210	19082	42	12977.58

Status: Tags: Types:

Links to:

Key:

Value:

Property	Value
target	NCBI:2035
reference	pubmed:309223

Comment:

Coment 1

Query

Queries:

```
SELECT ?anno ?oligo ?user ?date
WHERE {
?anno abmaxis:subject ?oligo .
?anno abmaxis:author ?user .
?anno abmaxis:date ?date .}
```

Iteration 3: Inference and Rule Builder

- **Intent**
 - Provide reasoning support to assist in the maintenance of an efficient library design
 - Provide information regarding complex correlations between antibodies
- **Business Benefits/Results**
 - Ability to discover new information, improve access to information, potentially reduce time to insight
- **Technical Outcome**
 - Collection/generation of data is automated through implementation of rules

Inference Rules









Rule Selector:

Rule Name:

Rule Definition:

Description:

Semantic Web vs. Traditional Database

	Traditional Database		Semantic Web	
Generate new predicates not yet in model		Attributes must be column defined		Easily done with triple merges or rules
Link to external data records or web resources		Foreign keys don't work across databases		URI model guarantees connectivity anywhere
Mix semantics from more than one source		Requires a new table with controlled terms		Requires simply a URI to the new namespace
Detect contradictory or malformed statements		Requires extra process code		Can be automated using standard reasoners

Experiment Results – Feasibility

Goal	Status	Comments
Develop and implement flexible information model	✓	Knowledge on how to model/create an ontology is needed; Work with Ontology Modeler
Provide user-friendly UI	✓	Easily modifiable to support business needs – interface critical for proper use
Convert data to RDF	✓	Architecture is expandable to support other data formats

Experiment Results – Functionality

Goal	Status	Comments
Develop capability to search, browse and view	✓	Customized queries provides user with ability to organize/retrieve data not provided by existing UI
Provide ability to annotate electronic data	✓	Governance model needed to address potential issues [i.e. legality]
Provide capability to infer relationship	✓	Currently deductive reasoning in place, inductive reasoning can be implemented

Experiment Results – Business Value

Goal	Status	Comments
Flexible data model	✓	Traditional database/web applications are not practical in this dynamic research area
Access to a centrally, integrated data source	✓	Able to update and annotate new data on an 'as-needed' basis
Useable Interface	✓	Customizable to support dynamic research needs regarding biologics

Conclusions

- Integration into dynamic area such as Biologics was possible
 - Enabled sharing and reuse of data (independent of type)
 - Facilitated the process of capturing knowledge via reasoners
 - Provided flexibility
 - Kept up with changes over the duration of the experiment
 - Integration of multiple data sources
 - Excel files
 - Centralized Library Resource
 - Extendable to other sources
- Not Enterprise ready for us
 - Not feasible to custom develop for each lab
 - Exploring availability of an application suite to provide proven capabilities with simple configuration

Thank you

Jaime Melendez

jaime_melendez@merck.com

Technology Innovation

Merck and Co., Inc.