



Semantic Technologies in Practice: Taking the Pain Out of Data Integration

June 7th

Semantic Technology Conference 2011
San Francisco, CA

Robert Stanley,
CEO, IO Informatics, Inc.

Core Methods

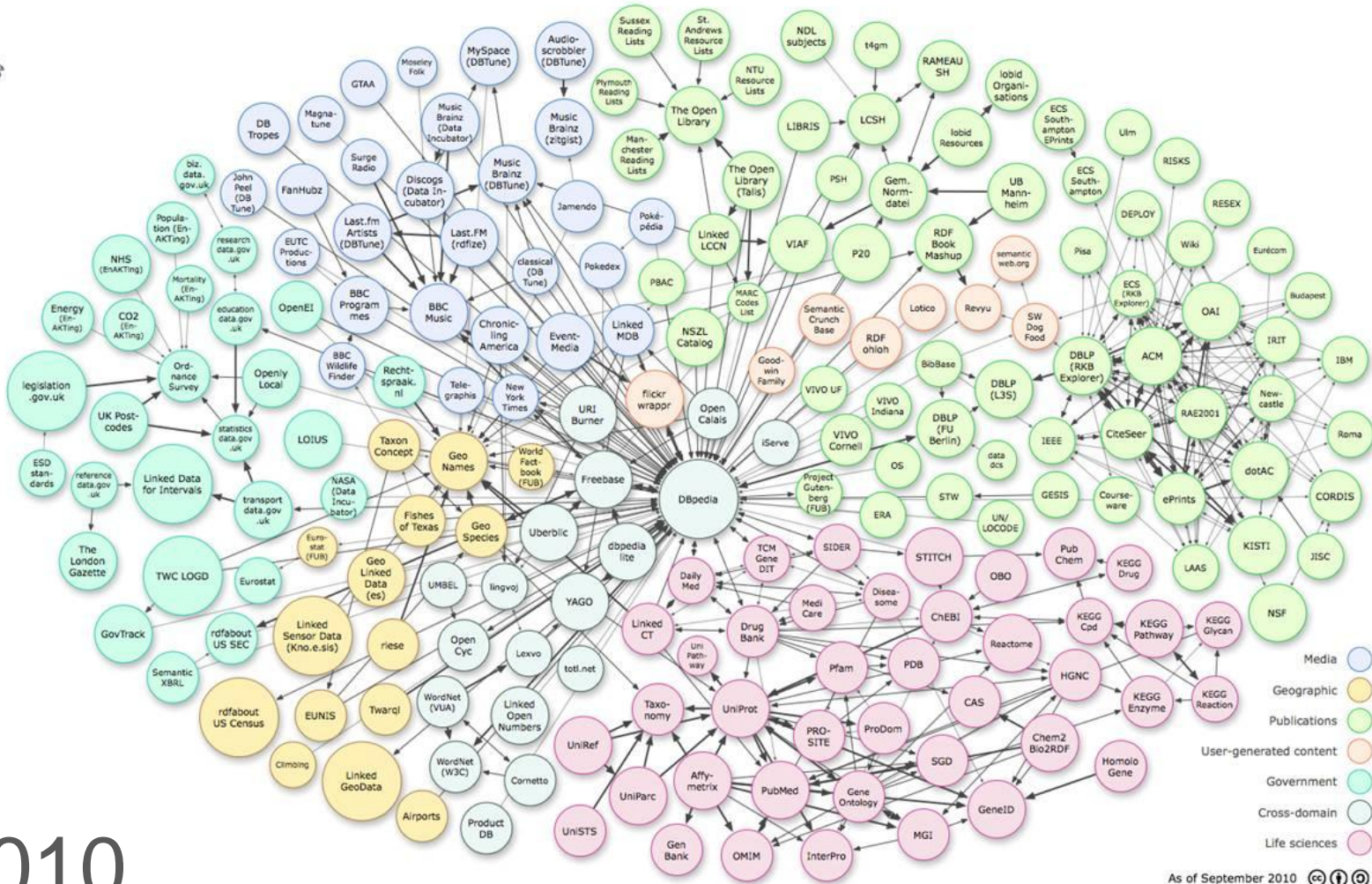
Core methods are maturing:

- Formal ontologies, application ontologies, W3C, NCBO, LOD and related resources
- Ontology alignment, inference, URIs, SPARQL endpoints, federation methods
- Scalability, security, support for transactional processing
- Expertise, training, larger projects (FDA, DOD, NASA, Pfizer, World Cup, Data.gov, etc.)



INFO

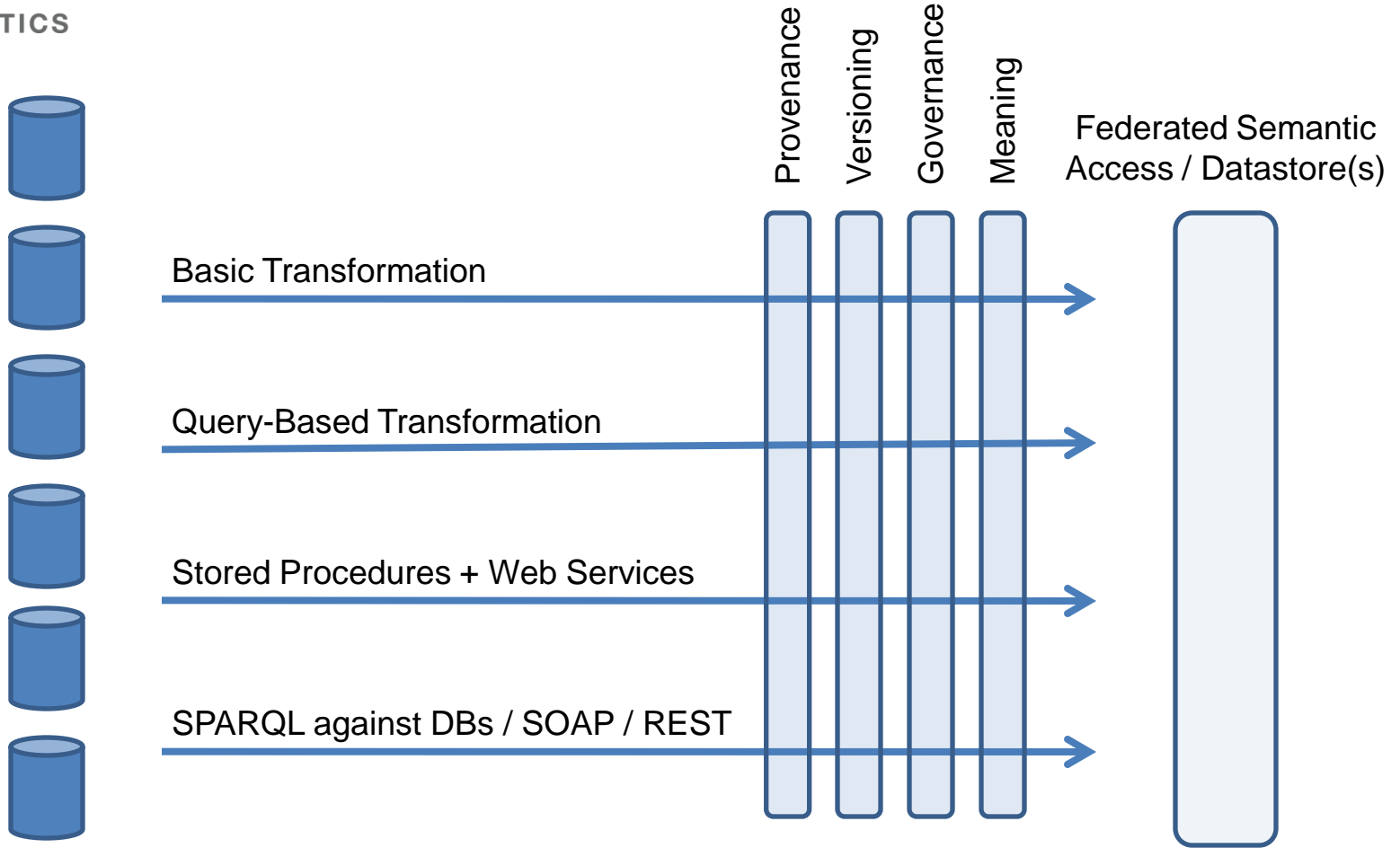
Exponentially Growing Resources



As of September 2010

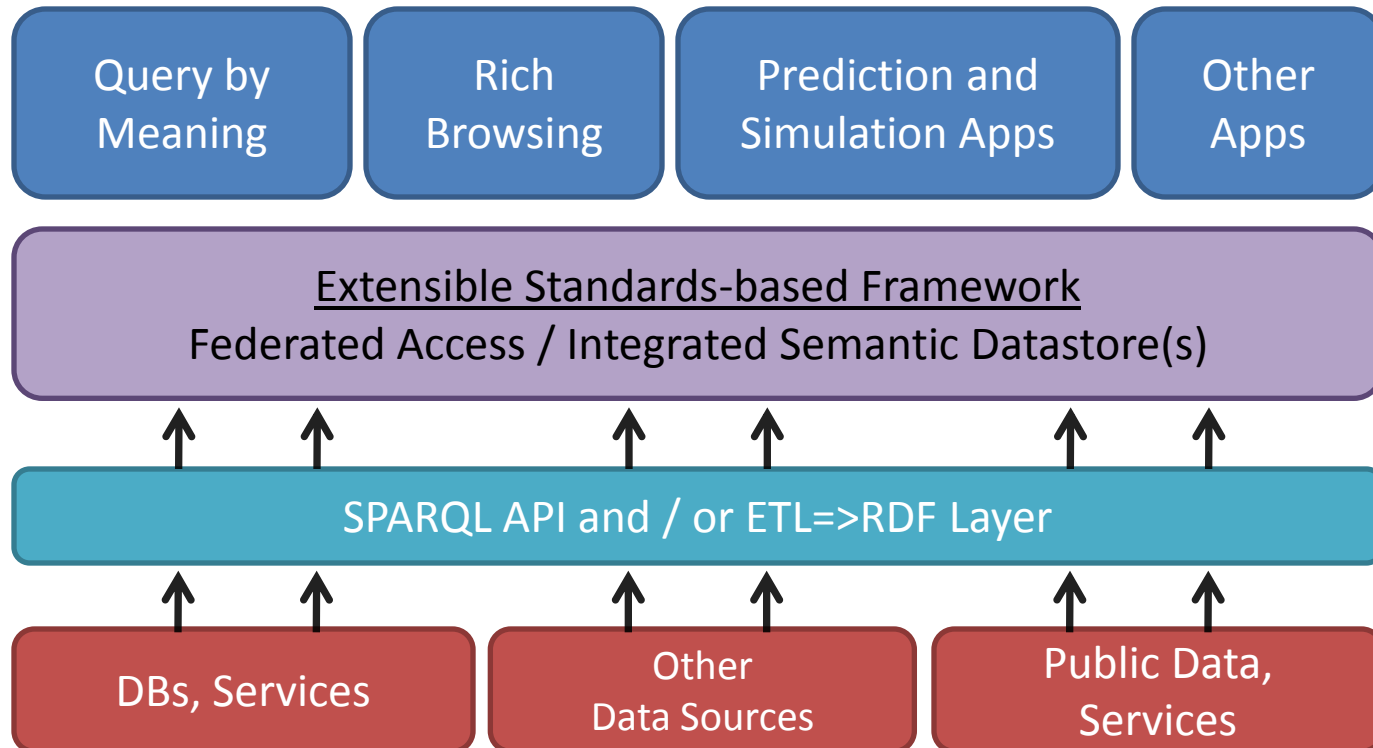
2010

Integration / Federation Options



Data Sources
RDBMS / REST / Web Services

RDF / SPARQL Solution Stack



... so what?

Quantitative Benefits

Reduced effort, time and cost to deliver and maintain value to consumers

- Orders of magnitude savings

Reduce communication overhead, design and deployment burden

- Data model is explicit, can be altered without refactoring

Reduced time to create and test integration

- Agile modeling, integration and testing

Extend to new data sources and applications

- “Building blocks” to add new data, create new applications

Qualitative Benefits

Growing access to resources

- Public data and ontology resources will make most expensive packaged DBs obsolete

R&D concepts can now be translated to immediate consumer benefits

- Previous “out of reach” integration and applications become practical

Emergent Properties

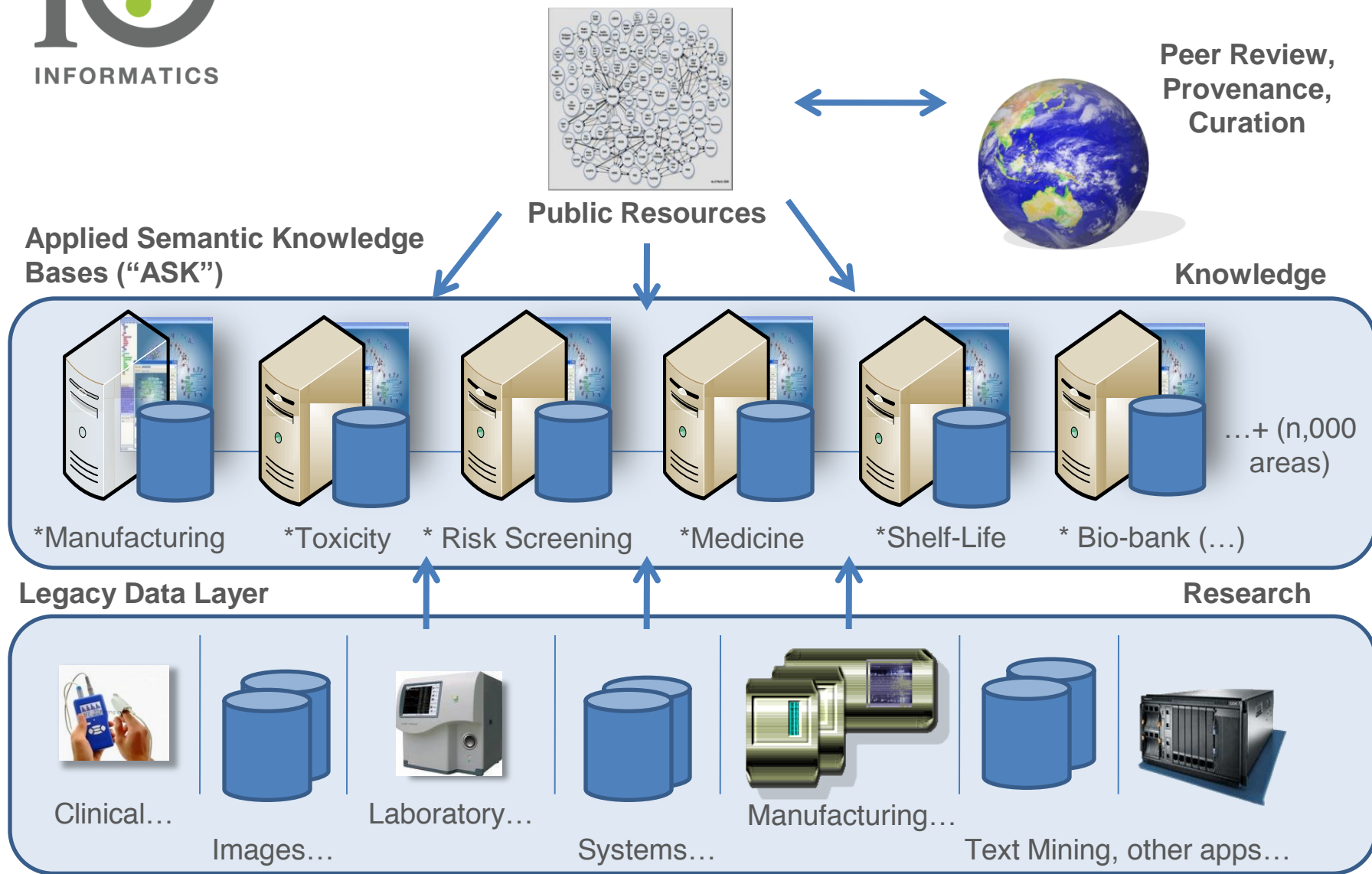
- Enrichment made easy!
- Supports inference, rich interrogation, serendipitous discovery



Business Examples

That's easy to say...

Let's Look at a Few Business Cases



Use Cases: Taking the Pain Away



- Manufacturing: Link data sources across imprecise connections to validate reports - in minutes rather than days



- Animal Safety: Knowledge Network for discovery, qualification and validation of cross-species biomarkers – reduces animal testing



- Personalized Medicine: Knowledge Network and screening application for personalized medicine – saves lives

Example Questions



- What data sources support this manufacturing report about product purity and shelf-life?



- What toxicity indicators are common to most animals?



- What combination of “biomarkers” can be used to detect risk of organ failure?



Pfizer: Reduce Time and Improve Reproducibility for Manufacturing Reports

Overview

- Integrate reporting system and experimental data sources
- Match purity analysis data with FDA reports for report verification
- Rank results based on multiple optional matches
- Web-based UI with parameterizable searches



Pfizer: Reduce Time and Improve Reproducibility for Manufacturing Reports

Challenges

- Lack of consistent and precise identifiers
- Use subject matter expert rather than IT staff for integration

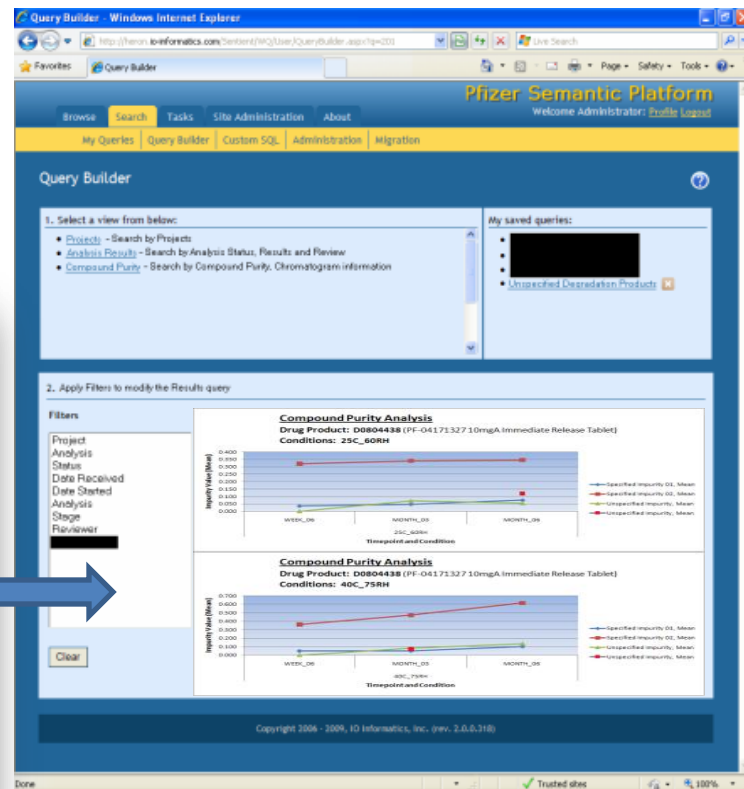
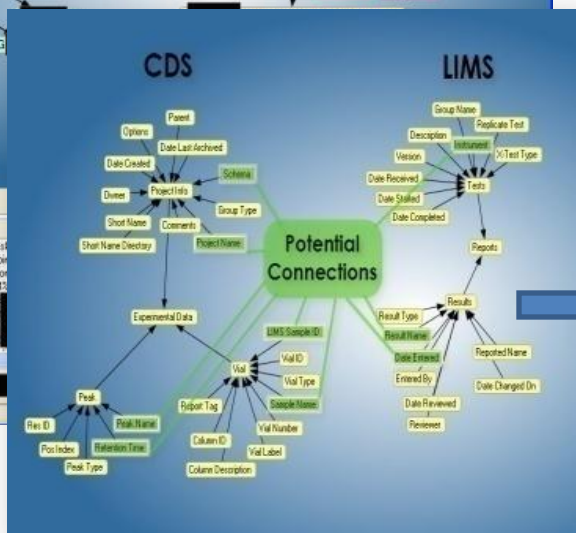
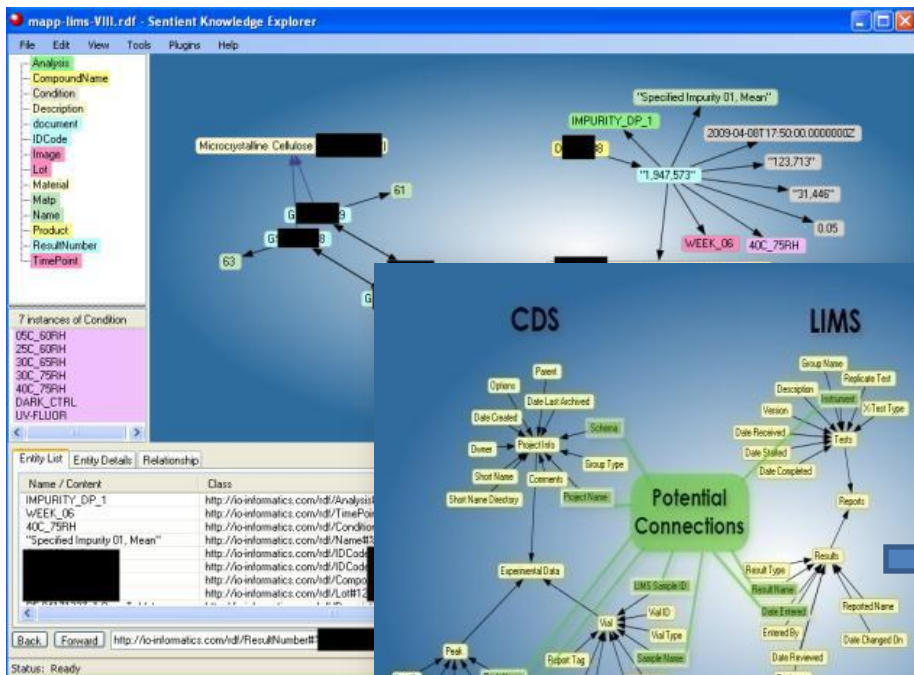


Pfizer: Reduce Time and Improve Reproducibility for Manufacturing Reports

Outcome

- Agile integration took significantly less time / effort than expected
- Web-based “Manufacturing Report Verification” application with best match ranking through parameterizable SPARQL queries
- Additional application for “Formulation Shelf-Life Prediction”

Manufacturing Report and Formulation Influence on Drug Stability



- Semantic integration of multiple data sources allows immediate web-based report verification and review for impact of “fillers” and compound formulation on stability of drugs.



Pfizer: Quantitative Benefits

- Project completion reduced from estimated 4 months to 4 weeks
- Report verification time reduced from 2 days to 1 hour
- Added formulation database in < 1 week for new formulation / shelf-life analysis application

Pfizer: Qualitative Benefits

Qualitative benefits

- Query array ranks results based on multiple imprecise identifiers
- Subject Matter Experts (SMEs) can lead on basic integrations and queries
 - Reduced reliance on IT staff, 1.5 FTE reduced to 1FTE
- New Application – Formulation and Shelf-Life Prediction
 - Integrated formulation data with existing purity and stability RDF
 - How do different formulations impact shelf-life?
 - Integration, testing and deployment completed in less than 1 week



FDA: Discovering Common Safety Indicators

Overview

- Create a “Knowledge Network” that can integrate changing data sources
- Surface patterns as queries to qualify, validate and apply cross-species indicators for safety / toxicity

FDA: Discovering Common Safety Indicators

Challenges

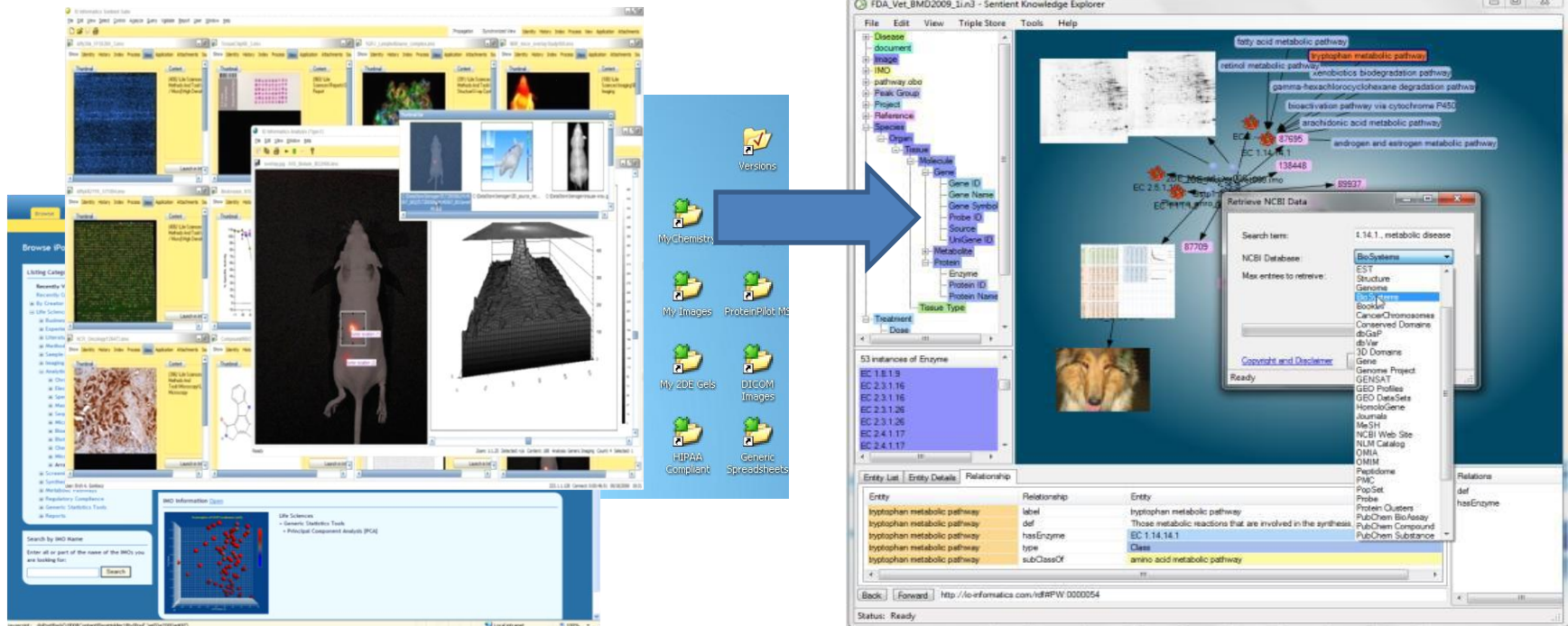
- Cost and time-efficient integration of changing experimental, correlation and published data
 - Multiple animal (i.e. mouse, porcine, horse, canine) and *in vitro* models (tissue, cell line)
 - Genetic and protein data
 - Image (pathology) data
 - Published data resources; e.g. NCBI Protein, NCBI Taxonomy, OMIM, PubMed, (LODD)



FDA: Discovering Common Safety Indicators

Outcome

- Subject Matter Experts create knowledge network with minimal assistance, can add new data
- Users traverse datasets, identify and qualify cross-species biomarkers, create patterns for validation



The screenshot displays the IO Informatics software interface, which is used for analyzing biological data across different animal species. The interface is divided into several sections:

- Left Panel:** A navigation pane with categories like 'Document', 'Image', 'Pathway', 'Project', 'Reference', and 'Species'. Below this is a 'Browse iFC' section with a 'Recently Viewed' list and a search box for iFC names.
- Center Panel:** A large workspace showing a 3D model of a rat, a 3D model of a brain, and a network diagram of metabolic pathways. The network diagram includes nodes for enzymes (EC 1.14.14.1, EC 2.8.1.1, EC 1.1.1.1, EC 1.1.1.1) and pathways such as 'fatty acid metabolic pathway', 'tryptophan metabolic pathway', 'retinol metabolic pathway', 'venobios biodegradation pathway', 'gamma-hexachlorocyclohexane degradation pathway', 'isochlorogenic acid metabolic pathway', 'arachidonic acid metabolic pathway', and 'androgen and estrogen metabolic pathway'. A 'Retrieve NCBI Data' window is open, showing search results for '134.1. metabolic disease' and a list of databases to search.
- Right Panel:** A 'Triple Store' window showing a list of instances of an enzyme (EC 1.14.14.1) and a table of relationships between entities.

The 'Triple Store' window shows the following relationships:

Entity	Relationship	Entity
tryptophan metabolic pathway	label	tryptophan metabolic pathway
tryptophan metabolic pathway	def	Those metabolic reactions that are involved in the synthesis
tryptophan metabolic pathway	hasEnzyme	EC:1.14.14.1
tryptophan metabolic pathway	type	Class
tryptophan metabolic pathway	subclassOf	amino acid metabolic pathway

- Genes, proteins, and other data (images, etc.) are analyzed across different animal species, for discovery of species-independent safety / toxicity biomarkers



FDA: Quantitative Benefits

- No initial estimate, initial integration completed in < 3 weeks FTE
- Ability to handle new data source structures with ease
 - *“... the hodgepodge of data is not only diverse, but also comes from a number of public and contracted sources.”*
- Ability for end users to integrate and analyze new data on the fly reduces response time from 2-3 days to about 2 hours

Reduced Animal Testing:

Animal testing costs the American public more than \$136 billion every year.



FDA: Qualitative Benefits

- Lowered barrier to entry makes the project possible
- Ability to share data across research domains, previously disconnected resources now available to all researchers
 - *“[The technology] has allowed us to transition to a more integrated approach.”*
- Discovery of common indicators for safety / toxicity will reduce need for animal testing. Initial results are being qualified and tested.



UBC: Knowledge Network for Organ Failure; “ASK” for Personalized Medicine

Overview

- Knowledge network for enrichment, visualization and qualification of patterns indicating risk of organ failure
- Web-based validation and deployment of “ASK” screening patterns indicating patients-at-risk



UBC: Knowledge Network for Organ Failure; “ASK” for Personalized Medicine

Requirements / Challenges

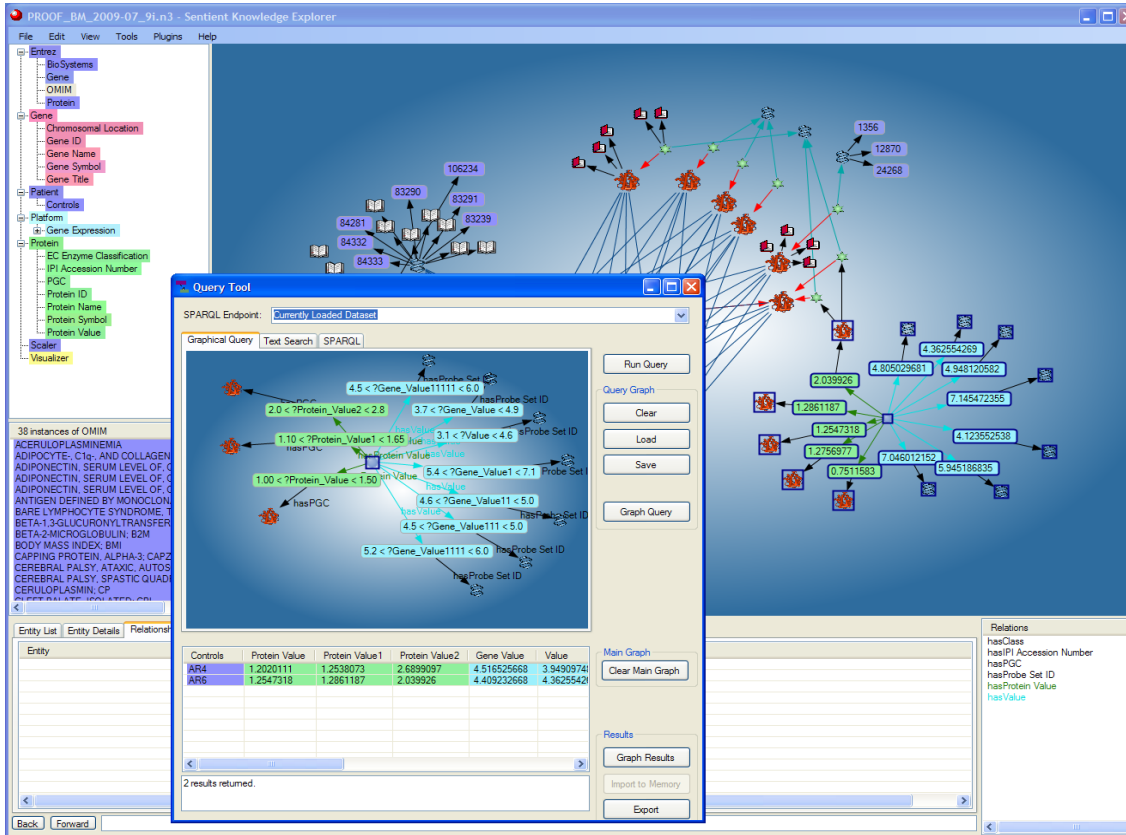
- Integrate, search, visualize multiple data sources
 - Genetic and protein data
 - Age, ethnicity, disease state, and clinical repositories
 - BioSystems, EntrezGene, GEO, HMDB, PubMed, LODD, (public resources)
- Create, test, apply patterns for patients-at-risk
 - Filter according to ranges, weights, inclusion / exclusion criteria
 - Report and visualize closeness of fit to pattern



UBC: Knowledge Network for Organ Failure; “ASK” for Personalized Medicine

Outcome

- Knowledge network for hypothesis generation and qualification, pattern capture and testing
- Applied semantic knowledge base (ASK) for clinical screening based on combined biomarker patterns



- Web-based Knowledge Application
- Applies patterns for predictive screening
- Weighing, scoring of results
- Bring “hits” back into Knowledge Network for validation of hypotheses and algorithms

- Screening of transplant patients for likelihood of transplant failure, based on combined biomarker patterns



UBC / PROOF: Quantitative Benefits

- Integration and analysis time reduced from about 2 years to about 8 months FTE equivalent
- Time to capture and apply patterns reduced from days to hours
- Knowledge base can be extended to include new public sources in hours (days / weeks with curation)
 - Expensive commercial database no longer needed due to ease of integrating public resources

UBC / PROOF: Qualitative Benefits

- Provenance, reference annotation, original data available for review
- Visual SPARQL presents queries as hypotheses
 - Make it possible for researchers to iteratively create, test and refine hypotheses
- Extended SPARQL delivers practically useful classifiers
 - Lowered barrier to move from research to application

“[The] ability to consume and intuitively represent a wide variety of data-types - from images to quantitative data - and more importantly, display that data in ways that make the significant features immediately obvious to our biologist end-users, has allowed us to move to a completely new level of data analysis....”



Recap – Core Technology Benefits

Semantic technologies deliver transformative quantitative and qualitative benefits

Agile: Semantic technologies reduce effort, time and cost to deliver and maintain integration-oriented databases and applications

Extensible: Ontologies and RDF provide building blocks for changing, growing integrations

Value-adds: Growing public resources are a fantastic catalyst

Transformative: Projects that were possible but impractical are now practical, even “painless” to achieve

Thank you!

Questions?



Email: rstanley@io-informatics.com

Website: www.io-informatics.com