ON THE ROAD TO PRODUCTION:
SEMANTIC INTEGRATION CASES INDICATE SUCCESSFUL ADOPTION TO IMPROVE KNOWLEDGE-BASED DECISIONS IN PHARMA AND HEALTHCARE

Erich Gombocz
VP & CSO, IO Informatics, Inc.

NCBO Webinar
Wednesday, April 17, 2013, 10:00 AM PDT
OUTLINE

- **State of Semantic Integration**
  - Adoption ongoing for years … - Why?
- **Challenges in Using LLD / LOD**
  - Approaches, Roadblocks & Solutions
- **Moving to Production**
  - Walk-through: How to Achieve Immediate ROI
- **Successful Scenarios**
  - Industry, Government & Clinical Examples
- **Live Demo**
- **Take Home**
  - Where we are, where we want to be
- **Consequences & Future Outlook**
- Acknowledgements
- References
STATE OF SEMANTIC INTEGRATION

WHAT HAS BEEN ACHIEVED & WHAT NEEDS TO IMPROVE
• RDF HAS EVOLVED AS ACCEPTED FRAMEWORK
  • DYNAMIC, EXTENSIBLE, INTEROPERABLE SOLUTIONS NEEDED FOR BIG DATA
  • NO NEED TO KNOW A PRIORI WHICH QUESTIONS TO ASK
• THE LOD CLOUD IS GROWING …
• SPARQL 1.1 IS W3C RECOMMENDATION (MARCH 21, 2013)
• LOTS OF POC, PILOTS

BUT

• TOO IDEALISTIC EXPECTATIONS:
  ***** LINKED (OPEN) DATA ≠ ***** COLLABORATIVE USABILITY
• DIVERGING DIRECTIONS:
  • DIFFERENT VOCABULARIES, REGISTRIES, OBJECTIVES, DESCRIPTORS
  • DIFFERENT PROVENANCE APPROACH (VoID, OpenPHACTS, Bio2RDF, BioDBCore, SADI, MIRIAM)
  • W3C HCLS TRIES TO RESOLVE THIS BY BUILDING CONSENT ON MAPPINGS
CHALLENGES IN USING LLD / LOD
APPROACHES, ROADBLOCKS & SOLUTIONS
NAVIGATING OBSTACLES

- ILL-DEFINED OBJECTIVES
- MISSING DOMAIN EXPERTISE: DATA RELATIONSHIP GUESSWORK
- INCONSISTENT NAMESPACE POLICIES
- USE OF INTERNAL, NON-DESCRIPTIVE ONTOLOGIES
- MISALIGNED PUBLIC, EXPERIMENTAL AND CORPORATE STANDARDS
- VERSIONING AND PROVENANCE ISSUES
- RELIABILITY FROM SERVICE-LEVEL TO URI PERSISTENCE
- CLINICAL INTEROPERABILITY (HL7, MedDRA, CDISC, MeSH, ICD9/10 …)
- MORE AND MORE “OPEN DATA” ARE CLOSED FOR COMMERCIAL USE
- SERIOUS FUNDING CONCERNS ABOUT GOVERNMENT-BACKED RESOURCES
APPROACH

RESOURCE ANALYSIS

• WHAT DO WE NEED TO ACCOMPLISH OBJECTIVES?
  • A SENSE WHAT WE WANT
  • BASICS (PROVENANCE, VERSIONING, HIGH INTERLINK QUALITY, PERSISTENCE)
  • GENERALLY APPLICABLE, QUICK & EASY SOLUTION

• FOCUS ON TRUE ‘★★★★★’ RESOURCES

• CREATE DYNAMIC “APPLICATIONS ONTOLOGY” FIRST

• USE EXISTING FORMAL ONTOLOGIES (OR PARTS OF) WHENEVER POSSIBLE
  • NCBO BioPortal

• HAVE INTEROPERABILITY FROM THE BEGINNING IN MIND
ACHIEVABLE?

YES, IT HAS BEEN DONE SUCCESSFULLY!
MOVING TO PRODUCTION

WALK-THROUGH: HOW TO ACHIEVE IMMEDIATE ROI
ROADMAP TO MAKE IT WORK

Map Data to RDF
- Namespace, URI Policy
- Entities vs. Literals, Data Types
- Scripted Transformations
- Provenance, Versioning, Attribution

Harmonize
- Concept alignment
- Vocabularies, Thesauri
- Ontology Merging

Refine Content
- Import only what’s needed
- Iterate via Visual SPARQL Queries
- Establish Classifier Patterns

Annotate & Add to Knowledge Base

Get Answers from Contextual Queries
<table>
<thead>
<tr>
<th>Ontology</th>
<th>Version</th>
<th>Release Date</th>
<th>Creator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass spectrometry</td>
<td>3.4.5.0</td>
<td>1/22/2013</td>
<td>PSI-MS Administrator</td>
<td>A structured controlled vocabulary for the annotation of mass spectrometry experiments. Developed by the HUPO Protein参考文献组.</td>
</tr>
<tr>
<td>Master Drug Data Base</td>
<td>2009_08_05</td>
<td>8/5/2009</td>
<td>Karen Eckert</td>
<td>Master Drug Data Base, 2009_08_05</td>
</tr>
<tr>
<td>Measurement Method Ontology</td>
<td>2.0</td>
<td>2/5/2013</td>
<td>Mary Shrimpton</td>
<td>The Measurement Method Ontology is designed to represent the variety of methods used to make qualitative and quantitative analyses in the life sciences.</td>
</tr>
<tr>
<td>Medicinal plant anatomy and development</td>
<td>unknown</td>
<td>6/9/2011</td>
<td>Medicinal_plant</td>
<td>The Ontology of medicinal plant anatomy and development of the Japanese medicinal flora.</td>
</tr>
<tr>
<td>Medical Diagnostic Categories Diagnosis Related</td>
<td>1.2</td>
<td>8/8/2012</td>
<td>K.M. Batho</td>
<td>Ontology of Medical Diagnostic Categories Diagnosis Related Groups</td>
</tr>
<tr>
<td>Medical Subject Headings (MeSH)</td>
<td>2012-2011_09</td>
<td>9/8/2011</td>
<td>Stuart Nelson, M.</td>
<td>Medical Subject Headings (MeSH) National Library of Medicine, 2011</td>
</tr>
<tr>
<td>Mental Functioning Ontology</td>
<td>alpha-2012-11-03</td>
<td>2/9/2013</td>
<td>Janani Hastings</td>
<td>The Mental Functioning Ontology is an ontology for mental functioning, including mental processes such as cognition and traits such as intelligence, and related diseases.</td>
</tr>
<tr>
<td>MetACT Ontology</td>
<td>11/8/2012</td>
<td>6/12/2012</td>
<td>Clarissa Rodriguez</td>
<td>An ontology to semantically describe and analyze clinical trials</td>
</tr>
<tr>
<td>Metagenome/Microbes Environmental Ontology</td>
<td>0.2</td>
<td>4/24/2012</td>
<td>MicrobeDB Editor</td>
<td>Ontology of environments, particularly focused on microbes.</td>
</tr>
<tr>
<td>Microarray Data Expression Ontology</td>
<td>1.3.1.1</td>
<td>2/9/2007</td>
<td>MGED</td>
<td>Concepts, definitions, terms, and resources for standardized description of microarray experimental in support of MAE v1.</td>
</tr>
<tr>
<td>Molecule Culture Collection Vocabulary</td>
<td>Version 0.8 beta</td>
<td>1/31/2013</td>
<td>Shuhui Kawasaki</td>
<td>Structured controlled vocabulary for describing meta information of microbial culture collection maintained in biological research centers.</td>
</tr>
<tr>
<td>Molecule Namebase Ontology</td>
<td>Version 0.2 alpha</td>
<td>2/18/2013</td>
<td>Shuhui Kawasaki</td>
<td>An ontology for describing microbial phenotypes.</td>
</tr>
<tr>
<td>miRNA Ontology</td>
<td>1.5</td>
<td>1/15/2013</td>
<td>Vicki D’Azzo</td>
<td>An application ontology for miRNAs.</td>
</tr>
<tr>
<td>Minimal anatomical terminology</td>
<td>1.1</td>
<td>12/6/2012</td>
<td>EMAP Administrator</td>
<td>Minimal set of terms for anatomy.</td>
</tr>
<tr>
<td>Mouse anatomy and development</td>
<td>1.9</td>
<td>12/6/2012</td>
<td>Michael Toussaint</td>
<td>A Mouse Ontology under construction designed to function of mouse genetic elements. The terms defined in MEGO are used to annotate shape and physiology.</td>
</tr>
<tr>
<td>Mouse Pathology</td>
<td>unknown</td>
<td>12/6/2012</td>
<td>Mouse_pathology</td>
<td>Mouse pathology ontology.</td>
</tr>
<tr>
<td>Multiple alignment</td>
<td>unknown</td>
<td>12/14/2010</td>
<td>Mac Administrator</td>
<td>An ontology for data retrieval and exchange in the fields of multiple DNA/RNA alignment, protein sequence and protein structure alignment.</td>
</tr>
<tr>
<td>NanoFib Protein Ontology</td>
<td>2011-12-08</td>
<td>12/6/2011</td>
<td>Nathan Baker</td>
<td>An ontology that represents the basic knowledge of physical, chemical and functional characteristics of nanotechnology used in cancer diagnosis and therapy.</td>
</tr>
<tr>
<td>Natural Products Ontology</td>
<td>unknown</td>
<td>2/8/2012</td>
<td>Risa Takashima</td>
<td>An ontology for describing biological activities of natural products.</td>
</tr>
<tr>
<td>NCBI organism classification</td>
<td>1.2</td>
<td>12/12/2011</td>
<td>NCBI Information</td>
<td>A hierarchical classification of living organisms and associated taxonomic terms for their description contained within the context of the databases.</td>
</tr>
<tr>
<td>NCI Cancer Ontology</td>
<td>12.04</td>
<td>4/30/2012</td>
<td>NCI Cancer</td>
<td>A vocabulary for clinical care, translational and basic research, and public information and administrative activities.</td>
</tr>
<tr>
<td>NCIThesaurus</td>
<td>12.04</td>
<td>4/30/2012</td>
<td>NCI Cancer</td>
<td>A vocabulary for clinical care, translational and basic research, and public information and administrative activities.</td>
</tr>
<tr>
<td>Neglected Tropical Disease Ontology (NTDO)</td>
<td>Rev 368</td>
<td>9/11/2012</td>
<td>Ripsi Santana</td>
<td>The Neglected Tropical Diseases Ontology (NTDO) aims at representing classes and relations to a specific set of diseases which persist in exactly the physical-psychosocial domain.</td>
</tr>
<tr>
<td>Neomark Oral Cancer Ontology</td>
<td>4.1</td>
<td>3/24/2011</td>
<td>Darío Salvi</td>
<td>Oral Cancer Ontology that describes the medical information necessary for early detection of the oral cancer recurrence extracted from the Neomark Project.</td>
</tr>
<tr>
<td>Neomark Oral Cancer-Certified Ontology</td>
<td>3.1</td>
<td>5/17/2012</td>
<td>Maria Ortega</td>
<td>Ontology that describes the medical information necessary for early detection of the oral cancer recurrence extracted from the Neomark Project.</td>
</tr>
</tbody>
</table>
| Neural Electro Magnetic Ontologies (NEMO) | v2.98 | 7/2/2012 | Gwen Finshoff     | Neural Electro Magnetic Ontologies (NEMO) describe classes of event-related brain potentials (ERP) and their properties, including spatial, temporal, and functional cognition.
Semantically Integrated Biological Networks are Leading to Actionable Knowledge
PHARMA SUCCESS STORIES
EXAMPLE 1

EXCIPIENT INFLUENCE ON DRUG PURITY, STABILITY & EFFICACY
RESULT: IMMEDIATE REPORT VERIFICATION AND MANUFACTURING BASED ON EFFECT OF COMPOUND FORMULATION ON DRUG STABILITY AND PURITY
EXAMPLE 2

PRE-CLINICAL TOXICITY RISK ASSESSMENT & CATEGORIZATION

NEED:
BIOMARKER-BASED TOXICITY ASSESSMENT AND CATEGORIZATION

CONVENTIONAL APPROACH:
3 YRS. OF EXPERIMENTS, STATISTICAL CORRELATIONS

SEMANTIC INTEGRATION:
LOD-BASED SYSTEMS-BIOLOGICAL QUALIFICATION OF EXPERIMENTAL PHARMACOGENOMIC CORRELATIONS
TOXICITY CLASSIFICATION
IDENTIFICATION OF TYPES OF TOXICITY (NIST ATP)

RESULT: FUNCTIONALLY QUALIFIED MULTI-MODAL (GENES, METABOLITES) BIOMARKERS TO IDENTIFY AND CLASSIFY TOXICITY
EXAMPLE 3

CLINICAL TRIAL DRUG SAFETY AND ADVERSE EFFECT ASSESSMENT
MULTI-NATIONAL PHARMA*

PROJECT SCOPE

- ~14,000 DOCUMENTS (~8GB) WITH
- ~1.2 BILLION SPREADSHEET CELLS CONTAINING DATA
- DATA CURATION REQUIREMENTS NOT OBVIOUS AT START

RESULT:

- INITIAL SEMANTIC KNOWLEDGE BASE
  - ~780 M Triples
- PARAMETERIZABLE SPARQL
- DRAG-&-DROP QUERY BUILDER
EXAMPLE 1

FDA VET (CVM) SPECIES-INDEPENDENT BIOMARKER INITIATIVE

NEED:
INTEGRATION OF GENOMICS, PROTEOMICS, IMAGING ENDPOINTS, ANIMAL DATA FOR PROJECT TO REDUCE ANIMAL TESTING

SEMANTIC INTEGRATION:
COMBINE EXPERIMENTAL AND ANIMAL DATA WITH PUBLIC RESOURCES INTO A COMPREHENSIVE MULTI-SPECIES BIOLOGICAL KNOWLEDGE BASE
CROSS-SPECIES BIOMARKERS REDUCING ANIMAL TESTING

RESULT: SEMANTIC INTEGRATION (LARGE ANIMALS TO SMALL ANIMALS TO CELL CULTURE) TO DISCOVER CROSS-SPECIES BIOMARKERS APPLICABLE TO HUMAN ADVERSE EVENTS AND DISEASES
EXAMPLE 2

NARMS MICROBIAL PATHOGEN KNOWLEDGEBASE

NEED:
INTEGRATION OF ICTV, MIST, PATRIC, PATHWAY RESOURCES AND EXPERIMENTAL DATA FOR QUICK PATHOGEN IDENTIFICATION

SEMANTIC INTEGRATION:
MICROBIAL PATHOGEN KB & APPLIED SEMANTIC KNOWLEDGEBASE (ASK)
MICROBIAL KNOWLEDGEBASE

RESULT: ACTIONABLE INTEGRATION ACROSS MULTIPLE RESOURCES FOR PATHOGEN DETECTION
CLINICAL USE CASES
EXAMPLE 1

COMPARATIVE EFFECTIVENESS OF COMBINATION TREATMENTS

NEED:
EVALUATION AND COMPARISON OF EFFECTIVENESS OF COMBINATION THERAPIES FOR PROSTATE CANCER

SEMANTIC INTEGRATION:
INTEGRATION OF 4 DIFFERENT GENOMICS PLATFORMS AND PROTEOMICS FOR BIOMARKER-BASED TREATMENT ASSESSMENT
RESULT: EFFECTIVENESS COMPARISON OF DIFFERENT COMBINATION TREATMENTS BASED ON MULTI-PLATFORM GENOMIC AND PROTEOMIC MARKER PROFILES AND PATIENT MATCH
EXAMPLE 2

ALERTS ON PRE-SYMPTOMATIC TRANSPLANT REJECTION

NEED:
BIOMARKER-BASED SCREENING OF HEART AND KIDNEY TRANSPLANT PATIENTS TO DETERMINE RISK OF ORGAN REJECTION

SEMANTIC INTEGRATION:
INTEGRATION OF GENOMIC AND PROTEOMIC BIOMARKERS AND CLINICAL ASSAYS FOR DECISION SUPPORT
Capture & Qualify Risk Patterns

Combinatorial Biomarkers for Organ Failure

Result: Integrated Knowledgebase for Combinatorial Marker-Based Screening of Transplant Patients for Likelihood of Organ Failure
DECISION SUPPORT
ASK FOR PRECISION MEDICINE

RESULT: SCORE-BASED RECOMMENDATION FOR IMMUNE SUPPRESSION THERAPY AT LIKELIHOOD OF REJECTION, AND PHYSICIAN ALERTING
EXAMPLE 3

PLAQUE STABILITY ASSESSMENT IN ARTERIOSCLEROSIS

NEED:
BETTER DETECTION OF EROSION OF CORONARY PLAQUES AND ONSET OF RUPTURE FOR ALERTS ON LIFE-THREATENING CONDITIONS

SEMANTIC INTEGRATION:
ESTABLISH GENOMIC, PROTEOMIC AND METABOLOMIC BIOMARKERS TO PREDICT PLAQUE RUPTURE
**CORONARY PLAQUE RUPTURE**
**RISK ASSESSMENT IN ACUTE ATHEROSCLEROSIS**

**RESULT:** INFLAMMATORY RESPONSE PRE-CURSOR MARKER CONFIRMS ONSET OF PLAQUE RUPTURE AND HELPS ACCELERATE RESPONSE IN ACUTE ATHEROSCLEROSIS
Example 4

COPD Biomarker-Based Clinical Decision Support System

Need:
COPD is a chronic progressive condition slowly damaging lung tissue; 2.2 B $ (2010, Canada). Biomarker-tests to detect and predict lung attacks

Semantic Integration:
Genomic and proteomic biomarkers associated with patient history and demographics to generate score-based recommendations
COPD Decision Support System
LIVE DEMO

VISUAL SPARQL QUERY ACROSS MULTIPLE RESOURCES
TAKE HOME

WE CAME A LONG WAY, BUT THE ROAD IS STILL LONG …
CONSEQUENCES, FUTURE OUTLOOK
THE PERFECT IS THE ENEMY OF THE GOOD & ECONOMIC IMPACT
**SEMANTIC TECHNOLOGY IS PRODUCTION-READY!**

- **Semantic Integration and Modeling in RDF/OWL** saves time, money, lives and helps discovering otherwise uncovered data inconsistencies.
- **Enriching experiments with LOD resources** facilitates better and faster qualification!
- "**Application ontologies**" are sufficient for most projects.
- **There is still room for improvement**
  - *Permanent URLs, better inter-linking and provenance for interoperability*.
- **Need to recognize socioeconomic benefits**
  - *Time and money saved should lead to new business models of private / academic partnerships to secure LOD resource funding in the future*.
ACKNOWLEDGEMENTS

COGENICS
Pat Hurban, Alan Higgins, Imran Shah, Hongkang Mei, Ed Lobenhofer

BOWLES CENTER FOR ALCOHOL STUDIES / UNC
Fulton Crews

FDA VET
Haile Yancy, David Heller, Rudell Screven

FDA NARMS
Sherry Ayers

IO INFORMATICS
Andrea Splendiani, Jason Eshleman, Robert Stanley

SIB / Uniprot Consortium
Jerven Bolleman

Wikimedia Foundation
Anja Jentsch

BIO2RDF II
Michel Dumontier

BMIR / NCBO Stanford
Mark Musen, Trish Whetzel

W3C
HCLS / Pharmacogenomics SIG

Support for Toxicity Studies
NIST ATP #70NANB2H3009
NIAAA #HHSN28120051008C
REFERENCES


3) **BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.** Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. Nucleic Acids Res. 2011; 39 (Web Server issue): W541-5


6) **VoID Vocabulary of Interlinked Datasets.** Cyganiak R, Zhao J, Alexander K, Hausenblas M. DERI, W3C note 6-Mar-2011

7) **PROV-O: The PROV Ontology.** W3C Candidate Recommendation 11- Dec-2012


13) **Ontology-Based Querying with Bio2RDF’s Linked Open Data.** Callahan A, Cruz-Toledo J, Dumontier M. 2013. Journal of Biomedical Semantics; in press.
DISCUSSION

THANK YOU!
egombocz@io-informatics.com