Business Semantics as an Interface between Enterprise Information Management and the Web of Data:

A Case Study in the Flemish Public Administration

Christophe Debruyne and Pieter De Leenheer eBISS, July 2012



Vrije Universiteit Brussel





- Give an understanding of Enterprise Information Management for Semantic Interoperability (by means of a case)
- Understand the role of conceptual modeling and the tension field between
 - Reusability and usefulness of models
 - Types of semantic interoperability
- Identify requirements for methods for EIM
- Present a method (and tool) for EIM: Business Semantics Management

Introduction

- Christophe Debruyne
 - Vrije Universiteit Brussel, STARLab
 - Method, Tools and Application of Ontologies
 - MSc in Computer Science @ VUB in 2009
- Dr. Pieter De Leenheer
 - Vrije Universiteit Amsterdam, Business, Web & Media
 - Service science
 - Collibra NV/SA (<u>http://www.collibra.com</u>/)
 - MSc and PhD in Computer Science @ VUB in 2004 and 2009 resp.

Outline

- Context FRIS and CERIF
 - The need for enterprise information management and/ for semantic interoperability
- Terminology and pinpointing the challenges
 - Requirements for tackling those challenges
- Business Semantics Management
 - Framework and Collaborative Method
- Hands-on BSM

Part I: Context

CERIF

- Common European Research Information Format (CERIF)
- Recommendation to EU-members for the storage and exchange of current research information.
- Aim: greatly facilitate the **reporting** process
- Created by euroCRIS by means of Entity-Relationship diagrams
 - To cope with multiple languages
 - For flexibility
- Core concept: Researcher, Research Project, Research Group, Equipment, Publication, etc.





Flanders Research Information Space

http://www.researchportal.be/

- Boost innovation through
 - aggregating data and make it publicly available
 - 19683 projects
 - 1976 organizations
 - 13982 researchers
 - ...
 - multiple parties deliver
- Enterprise Information Management
- Linked Data Approaches



Combining Economy, Science and Innovation for a better society Copyright © 2008 Department of Economy, Science and Innovation, all rights reserved Solution provided by Atire 2008, all rights reserved Terms of use | About your privacy

Flemish Government Department of Economy, Science and Innovation Koning Albert II-laan 35 bus 10 B-1030 Brussel - Belgium E-mail:<u>fris@vlaanderen.be</u>

FRIS Business Drivers

- **Research institutions**: a standard semantic framework for reporting on research activities and results.
- **Policy makers**: accurate and timely overview of research activities and results to improve innovation.
- Funding agencies: identify challenges in research or opportunities for exploitation.
- **Researchers**: avoid wasting valuable time collecting the same data over and over.

FRIS





Linked Data and FRIS



Linked Data Web is another Story

- Many communities heterogenous goals
- Open Innovation paradigm: boost change through sharing
- First unlock data, then think about possible applications
 - Lightweight loosely coupled vocabularies
 - Generative technologies such as RDF, HTTP, SPARQL

Resource Description Framework

- RDF is not a language, but a model
- RDF is a W3C recommendation
- RDF is designed to be <u>read by computers</u>
- RDF is for describing resources on the Web
- RDF uses URIs to identify and reference resources on the Web



SPARQL

SPARQL Protocol and RDF Query Language

• is an RDF query language

```
PREFIX ex: < http://starlab.vub.ac.be/example# >
SELECT ?title
WHERE {
  ?thingy ex:title ?title.
}
                                              ?title
                                            STARLab
```

FRIS

- To populate the FRIS portal with all information provided by the delivered CERIF files and other heterogeneous sources, needed are:
 - Consensus amongst the involved parties on a common conceptual model for CERIF and the different classifications (inside that semantic layer);
 - An easy, repeatable process for validating and integrating the data form those sources;
 - Make available the information in a generic way on the Web on which third parties can develop services as demonstrated by other Linked Data initiatives.

Part II: Some terminology & identification of challenges

Enterprise Information Management and the Web of Data

- Knowledge management and conceptual modeling are important activities for both Enterprise Information Management (EIM) and the Web of Data.
- EIM (Top-down)
 - Satisfying IT needs emerging from organization's requirements
- The Web of Data (Bottom-up)
 - Provide structured data for (third party) services

Remember, FRIS aims at both!

Semantic Interoperability

- Semantic interoperability is defined as "the ability of two or more *autonomously* developed and maintained information systems (IS) to communicate data and to interpret the information in the data that has been communicated in a meaningful manner"
- When a need for semantic interoperability rises, the community of stakeholders formulates these in semantic interoperability requirements.

Closed vs. Open Systems

• Closed = developed within, and for, 1 organization

Semantic Interoperability NOT required

Open = developed for deployment on internet

Semantic Interoperability ENABLED by a shared ontology: Collaborative Ontology Engineering > COMMUNITY!

Ontologies

- The formal semantics of a (computer-based) system is the correspondence between this system and some real world as perceived by humans and usually given by a formal mapping of the system's symbols.
- As the real world is not accessible inside a computer, the world needs to be replaced by an agreed conceptualization if we want to store and reason about semantics. Semantics are often stored in the shape of a formal (mathematical) construct. The resulting artifact is what we call an ontology.

Other definition of ontology: a computer-based, shared, agreed formal conceptualization is known as an ontology.



The study of the categories of *things* that exist or may exist in some domain.

What concepts do exist?

How are concepts related to each other ?

How are concepts subdivided according to differences and similarities ?

Ontologies in Computer Science

- In summary: Semantics = Agreed Meaning
 - Links symbols in autonomously developed systems to shared reality
 - Agreed among humans as cognitive agents
 - Stored in ontologies
 - key technology for interoperability (SemWeb)
 - ontologies ≠ data models, but provide annotations for them
 - support both human- and system-based reasoning

Ontologies in Computer Science

- The problem is not so much what ontologies in computer science are, but how ontologies come to be.
- An ontology is the result of a series of interaction leading to agreements to a better approximation of a communities perceived reality, often for a specific goal.
- This goal is defined by the community's semantic interoperability requirements

Challenges

- Ontologies contain references to the instances used in the application or application domain, and domain rules.
 - Domain rules typically contain constraints of identity, cardinality, mandatoriness, etc. and thus restrict the semantics (i.e. interpretation) in a specific conceptualization of a particular application domain.
 - Depending on the semantic interoperability requirements, domain rules can be crucial. E.g., conceptual reference structures.
 - Providing more rules, however, also reduces the generality of these ontologies.

Challenges

- Tension field Web of Data >< Enterprise Information Management (EIM)
- Describing existing (legacy) data can be done with lightweight ontologies.
- However, as more business rules are needed to ensure proper business within the community of stakeholders, EIM will be applied to capture the requirements on how and under what conditions data will be exchanged.

Challenges

- The Web of Data and EIM are thus residing in two different business domains and have different business drivers.
- Bottom-up (Web Of Data) vs. Top-down (EIM)
- For EIM: vocabulary management is central

Requirements for a Method

- Community involvement
- Learn from database modeling methods and techniques:
 - Technology matures
 - Analyzing natural language discourse
 - Employing legacy data, output reports, interviews, etc.
 - Lift data models into ontologies, remove applicationspecific context

Part III: Business Semantics Management

Six principles of BSM

- ICT Democracy: ontology should be defined by its community
- **Emergence**: semantic interoperability requirements emerge autonomously from community evolution processes
- **Co-evolution**: ontology evolution processes are driven by the changing semantic interoperability requirements
- **Perspective Rendering**: ontology evolution processes must reflect the different stakeholder perspectives
- **Perspective Unification**: relevant parts of the various stakeholder perspectives serve as input for the unified perspective
- Validation: validating ontology against these perspectives

A Brief History ...

- •1995. VUB STARLab was founded
- •1999. DOGMA Ontology Engineering Framework
- •2006. DOGMA-MESS
 - Meaning Evolution Support System
 - A method built on top of DOGMA
- •2008. Business Semantics Management (BSM) (DOGMA-MESS revisited)
- •2008. Collibra was founded
- •2010. Research in social processes in OE

- Fact-oriented. Communication of elementary fact-types by analyzing natural language discourse. Fact-types are "generalizations" of facts.
 - [Person] knows [Person] is a fact-type. [Christophe] knows [Pieter] a fact.

Different from frameoriented approaches!

- Elementary means fact-types can not be broken down (atomic)
- Lexon Base. A vast base of plausible binary fact-types called lexons.



Lexon paths



- Commitments are selections of the lexon base with constraints to represent the domain
 - Community commitments contain the selection of lexons and constraints the community is supposed to commit to. He or she engages in at least adhering to those agreed upon facts and constraints.

• Represents the ontology!

• Application commitments furthermore contain annotations of the application symbols



"Double articulation principle"

- Commitments are selections of the lexon base with constraints to represent the domain
 - Community commitments contain the selection of lexons and constraints the community is supposed to commit to. He or she engages in at least adhering to those agreed upon facts and constraints.
 - Application commitments furthermore contain annotations of the application symbols

- Importance of domain rules.
- E.g., identification. Unique, total, and identifying set of attributes





Modeling communities in BSM

- Semantic communities
- Body of shared meaning
- Speech communities
- Vocabularies



- Vocabularies are part of
 Speech Communities.
 Speech communities are part of Semantic communities.
- All information (formal and informal) stored in this system is called a *glossary*.

Semantic Reconciliation

- Semantic Reconciliation. Business semantics are modeled by extracting, refining, articulating and consolidating fact-types from existing sources.
 - Results in a number of consolidated language-neutral semantic patterns (community commitments) that are articulated with informal meaning descriptions
 - These patterns are reusable for constructing various semantic applications.





- Scope sets out the scoped terms that are actually needed to establish semantic interoperability
 - Distinction between an IT- or IS-context and a business context. They imply different kinds of threats
 - Involve the relevant stakeholders in this process and assign them with appropriate roles and responsibilities.







- During this activity, every scoped term is syntactically defined.
 - In the CERIF Project Community
 - CFProject executed by / executes CFOrganiza Million
 - CFPerson having / of Person_Name

•

- CFPerson having / of CFPersonAddress
- CFPersonAddress of / used in CFAddress
- EACH CFPerson having EXACTLY ONE Person_Name

kijkt naar DOGMA METHODE DOCUMENT MET WIJN VOORBEELD INTERACTIVITEIT VOORZIEN



- During this activity, fact types that were created during the Create activity are refined so they are understandable to both business and technology
- CURE: Correct, Useful, Reusable, Elegant
- Objectification
- Capture missing link



Business Ser	mantics Gloss	sary		
	Business Semantics Glossary			Oministrator Openhosard Clossary Settings
Roles S Tas	ks 🛱 Tools ** Workflows			Q. Search
Attributes N	esearch Information Space > CERIF > Project			Information Status: Candidate Concept Type: Object Type Steward: Pieter De Leenheer 37.5%
Definitions	Planned set of interrelated tasks to be executively within certain cost and other limitations.	ted over a fixed period and	2	Attributes
Examples	Large Hadron Collider		2	Add Attribute
				Relations
Fact Types	 Project has Title Project funded by Funding Programme Project executed by Organisation Project has Budget Project has Duration Project ends on Date Project starts on Date Project described by Discipline Code 	(Candidate) ② ○, ○, (Candidate) ③ ○, ○,		Taxonomy
General Rule Set	Project described by at most 3 Discipline Code. Project executed by at least 1 Organisation. Project ends on exactly 1 Date. Project has at least 1 Title. Project starts on exactly 1 Date.	1 2 1 2 1 2 1 2 1 2		Stakeholders Felix Van De Maele Administrator Ceert Van Grootel



- Create informal meaning descriptions as extra documentation.
- Include definitions and examples.
- serve as anchoring points when stakeholders have used different terms for the same concepts (i.e., detecting synonyms).
- Where available, descriptions already existing can be used (e.g., the euroCRIS website on CERIF) to speed up the process and facilitate reuse.

A Project	± ∕ ₽	
Flanders Re	esearch Information Space > CERIF > Project	
Attributes Na	ames Categorization Scheme	
Attributes Na	ames Categorization Scheme	
Attributes Na	ames Categorization Scheme	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Attributes Na Definitions	Ames Categorization Scheme Planned set of interrelated tasks to be executed over a fixed period and within certain cost and other limitations.	\$\$ 2
Attributes Na Definitions	Ames Categorization Scheme Planned set of interrelated tasks to be executed over a fixed period and within certain cost and other limitations.	



- A new version of the glossary is generated, which is a "flattened" version of the community commitment that is generated in a timely manner.
- The glossary is the product of semantic reconciliation and serves as a uniform technical specification to implement semantic applications.
- This glossary can be represented in many formats, such as UML, OWL, or XSD, serving a wide variety of applications.*

* The how is out of the scope of this lecture. However, feel free to contact us after the lecture for more information.

Semantic Application

- Semantic Application. Existing information sources and services are committed to a selection of semantic patterns.
 - I.e., creating application commitments
 - The existing data itself is not moved nor touched!





- Given an application context (such as a workflow or business artifact), relevant concepts are selected from the EIM for a particular application. It may be required to add additional application-specific concepts and constraints that could not be agreed upon on the community level.
- Anyone an idea of an example of application specific concepts and constraints?





- Information systems are improved using the selected concepts. Depending on the application context, this can be implemented in different ways.
- Concretely, this boils down to data transformation, validation, and governance services. For example:
 - Integrating two or more XML structures by defining XSLT transformations to a shared XSD-formatted EIM.
 - The EIM may also be used to convert relational databases into RDF triple stores.



Semantic Interoperability Between Actors

- FRIS Ecosystem. Funding agencies wish to form appropriate review boards.
- IWT: "bottleneck lies in defining varying review boards with no conflict of interest"



Optimize Review Cycle by Automating Error-prone Tasks

Experts who are a **party concerned** in one of the project proposals are excluded from participating in the colleges of experts. Next to this, experts who are not party concerned, but have an affiliation with a division/service/department identical to that of one of the applicants, **may not take part** in the assessment of the corresponding projects.



Towards Linked Data for FRIS

Modeling ontologies have several advantages

- Fact-types and constraints are easily verbalized
- No distinction between entities and relations (factoriented, rather than frame-oriented)
- Grounded in natural language
- Disadvantages?
 - Based on FOL --> Decidability
 - No everything can be modeled (e.g., procedures).





Towards Linked Data for FRIS

- Datasources are annotated with the ontology
- By means of of-the-shelve solutions for "triplifying" relational databases to RDF triples (cfr. RDB2RDF community)

SPARQL:

PREFIX db: <http://starpc14.vub.ac.be/2020/resource/>
PREFIX person: <http://starpc14.vub.ac.be/ontologies/person.owl#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX orgunit: <http://starpc14.vub.ac.be/ontologies/organization.owl#>
PREFIX v: <http://www.w3.org/2006/vcard/ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX project: <http://starpc14.vub.ac.be/ontologies/project.owl#>
PREFIX d2r: <http://starpc14.vub.ac.be/ontologies/project.owl#>
PREFIX d2r: <http://starpc14.vub.ac.be/ontologies/project.owl#>
PREFIX ads: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX map: <file:/usr/local/d2r-server-0.7/vub.starlab.gospl.glosses.ewicase/ewi.n3#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vocab: <http://starpc14.vub.ac.be:2020/vocab/resource/>
SELECT DISTINCT ?first_name ?last_name WHERE {
}

?person person:Person_having_First_Name ?first_name. ?person person:Person_having_Last_Name ?last_name. ?affiliation orgunit:Affiliation_of_Person ?person. ?affiliation orgunit:Affiliation_of_Organizational_Unit ?org. ?org orgunit:Organizational_Unit_has_Name "Informatica en Toegepaste Informatica"@nl.} LIMIT 10

LIMIT 10

Results: Browse

Go! Reset

SPARQL results:

first_name	last_name
"JEROEN"	"HOPPENBROUWERS"
"Sara"	"JANSSENS"
"Nina"	"LANDAU"
"FENG"	"LIU"
"Leen"	"RYMENANS"
"Eric"	"TANTER"
"Damien"	"TROG"
"ANTONIO"	"VALLEJOS VARGAS JORGE
"Tom"	"VAN CUTSEM"
"SVEN"	"VAN SEGBROECK"

\$

But that's not all ...

Business Semantics Management

- From a high-level perspective, three different kinds of data exchange exist within large organizations:
 - Exchange of knowledge between people;
 - Exchange of understanding between people and information systems;
 - Exchange of data between disparate information systems.
- We presented the application of BSM for Semantic Interoperability

Business Semantics Management

- In the context of this lecture, we limited ourselves to binary facttypes in the example
- BSM, however, adopted SBVR. SBVR is an OMG standard providing amongst others means for:
 - Including "named" instances
 - Modeling n-ary fact-types
 - Unary fact-types: [Proposal] is accepted, [Project] is running
 - Unary fact-types are useful for describing "dynamic" business rules



Conclusions

- Presented an application of BSM the EIM method within the Flemish Government for Semantic Interoperability and Linked Data
- The role of conceptual modeling and some tension fields
- A notion of fact-orientation

