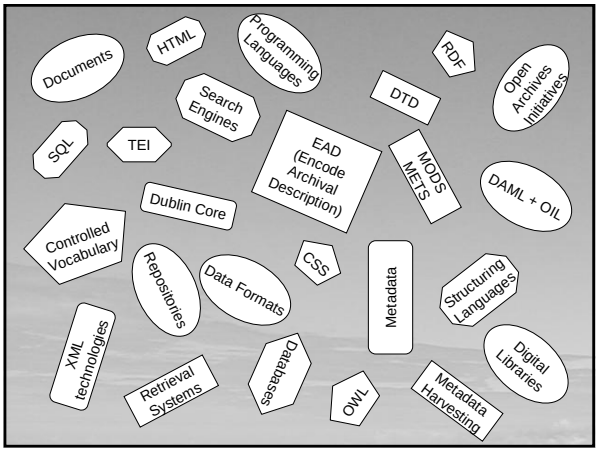
# Semantic Limitations of the Semantic Web

**Knowledge Management Symposium 2007**

Guillermo A. Oyarce, Ph.D.
Texas Center for Digital Knowledge
School of Library and Information Sciences
University of North Texas
oyga@unt.edu




## The Web

- Web resources are web pages, images, files, etc. transmitted using web-based protocols
- Each resource is identified by its location
- A Web locator is commonly known as URL (Universal Resource Locator)
- A URL is a type of URI, the Universal Resource Identifier
- Locator
  - Identifies a resource
  - Is a resource

## The Web

- The URL identifies the location of the resource, independent of its name
- The URN identifies the name of the resource, independent of its location
- A URI can be either a URL or a URN
- They are also treated as Web resources

http://www.w3.org/TR/uri-clarification/#contemporary

## The Web

- Metadata: Another type of resource
- Metadata
  - Data about the resource
  - Dublin Core
  - MARC (**MA**chine **R**eadable **C**ataloging)
- Types of metadata
  - Descriptive: Information in the resource
  - Structural: About format, nature of resource
  - Administrative: To manage, access, preserve resource
- Metadata Registries: Store information about different metadata sets

## The Semantic Web (SW)

- All resources have at least one URI
- A URI identifies only ONE resource
- The Resource Discovery Framework (RDF) encodes the URI and the metadata about a resource
- Documents for human consumption remain in some other repository
- Descriptive metadata is about the content in the document
- But the content in a metadata field is not always unambiguous and clear across sets
  - i.e. author versus creator, dates, types, etc.

http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

## The Semantic Web (SW)

- Software agents roam the SW to discover suitable resources to answer queries
- Data about documents, including metadata, are placed in discoverable resources (RDF)
  - Some standardization takes place through metadata
  - Different linguistic expressions of the same content can be represented by the same descriptor
  - The metadata representation of two documents with the same or very similar content may be the same
  - This creates semantic problems for software agents because they don't interact with the actual documents but with representations
  - Also, they interact not only with the registries of discoverable resources but with each other

## Semantic Descriptions

- Description of the resource
  - Location
  - namespace
- Description about the resource
  - Metadata
    - Management, service, descriptive
- Description about the information in the resource
  - Descriptive metadata
  - Controlled vocabulary
    - Taxonomy, thesaurus, ontology

## Inter-Agent Interaction

- Agents communicate
  - With each other
  - With data repositories
- Degrees of semantic communication
  - High level (human to human, or H2H)
    - Full text: narrative
    - Non-verbalized: implicit knowledge
    - Assumed: tacit knowledge
  - Low level (machine to machine, M2M)
    - Formal: hardwired in the process (hardware or software)

http://www.w3.org/TR/REC-rdf-syntax/

## Inter-Agent Interaction

- Agent-to-agent communication
  - What one says needs to be accurately interpreted by the other
  - Only explicit knowledge can be interpreted correctly
  - Flexibility and ambiguity = H2H
  - Formality and inflexibility = M2M, H2M
  - Semantic continuum:
    - from informal to formal
    - From implicit/tacit to explicit

## Semantics

- Human
  - Ambiguous
  - Dynamic
  - Unexpected
- Machine
  - Rigid
  - Formal
  - Explicit

## Discoverable Resources

- Location of resource
  - To know where to find it
- Description of type of resource
  - To know how to decode it
- Description about resource management
  - To know how to manage it
- Description about information in the resource
  - To know if the agent needs it

## Assumptions and Problems

- Assumptions
  - Common meaning among interacting agents
  - Common meanings across the SW
- Problems
  - No common vocabulary across the SW
  - No common metadata scheme
  - No common styles of implementations even among same standardized structures (i.e. using different policies to implement the same standard ontology language or thesaurus)

## Knowledge is slippery

- **Tacit** (if you don't know it we won't tell you)
- **Implicit** (understand-this)
- **Explicit** (know-this)
- **Declarative** (know-what, know-that)
- **Procedural** (know-how)
- **Strategic** (know-when, know-why)

## Inter-Agent Interaction

- Levels of interaction, communication
  - Signals (bits and bytes)
  - Protocols
- Formality
  - Explicit, machine-to-machine, unambiguous
- Interpretations
  - All assumptions have already been hardwired
  - Same language, intention, implementation and application

## The Semantic Web

**Primary Resources**

- Documents
- Data
- Information
- Knowledge

**Resource Types**

- Databases
- Knowledge base
- RDF
- Metadata repositories
- Information Retrieval systems
- Content management systems

## Languages

**HUMAN**

- Natural language
- Image and audio documents
- Formal but flexible
- Ambiguous
- Disambiguation is through common (tacit) knowledge

**MACHINE**

- Formal and rigid
- Policy implementation
- Hardwired use and interpretation
- Multilevel proliferation of languages
- One language level needs another level to assist in interpretation

## Web Languages for Machines

- RDF is a machine-to-machine language
- The RDF provides not only the structure but the language to identify resources
- Exists within its own structure, requires a repository of common understanding and a common view of the domain
- An RDF repository uses ontological markers to point at the resources
- Agents identify resources based on how they are described by the language

## Example

- Consider the following statement: "A Person identified by http://media.unt.edu:8080/ramgen/slisadminvideos/g._oyarcegreeting.rm, whose name is Guillermo Oyarce, whose email address is oyga@unt.edu, and whose title is Dr."

## Using XML to represent it

```
<?xml version="1.0"?>

<rdf:RDF xmlns:rdf="http://www.unt.edu/slis/people/faculty/oyarce.htm"
      xmlns:contact="http://www.unt.edu/slis/people/faculty/oyarce.htm">

 <contact:Person rdf:about="http://www.unt.edu/slis/contact.htm">
       <contact:fullName>Guillermo Oyarce</contact:fullName>
       <contact:mailbox rdf:resource="mailto:oyga@unt.edu"/>
       <contact:personalTitle>Dr.</contact:personalTitle>
 </contact:Person>

</rdf:RDF>
```

http://www.w3.org/TR/REC-rdf-syntax/

**NOTE:**
**What is represented is not the person, or even the information about the person but the resources that contains the information. The entry only describes the contents of the resources at the most basic semantic level.**

## Possible Variations

- Notice the emergence of possible problems. For <contact:fullName>Guillermo Oyarce </contact:fullName>
- Consider the following: <contact:Name>Guillermo Oyarce</contact:Name>
- Or <contact:firstName>Guillermo</contact:firstName> <contact:lastName>Oyarce</contact:lastName>

## Languages and Initiatives

**LANGUAGES**

- XML family (XML, XLS, XPATH)
  - http://www.w3.org/XML/
- OWL
  - http://www.w3.org/TR/owl-guide/
- DAML + OIL
  - http://www.daml.org/language/

**INITIATIVES**

- Open Archives Initiative (OAI) – Protocol for Metadata Harvesting (PMH)
  - http://www.openarchives.org/
- Encoded Archival Description (EAD)
  - http://www.loc.gov/ead/
- Text Encoding
  - http://www.tei-c.org/
- MODS (Metadata Objects Description Schema) and METS (Metadata Encoding and Transmission Standards)
  - http://www.loc.gov/standards/mods/
  - http://www.loc.gov/standards/mets/
- Dublin Core metadata
  - http://dublincore.org/

## Partial summary: Web Services

- A Web Service (WS) is a type of software in the Semantic Web (SW)
- A WS sends out agents to query data repositories to find specific data
- It can compile separate data and create dynamic documents
- Each partial component is a resource that may exist separately and independently
- RDF describes the resources and their locations
- RDF descriptions co-exist in data repositories
- RDF may include metadata and a controlled vocabulary
- The Web is becoming a large and complex database system

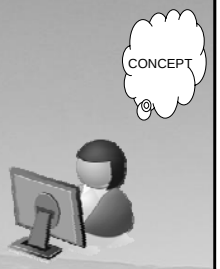
## Two Types of Resources

- Resources for machine utilization (location, metadata)
  - The SW infrastructure supports identification, finding, access and management of all sections of a document
- Resources for human consumption (documents)
  - While many services operate within the internal confines of the web, documents are the ultimate objective

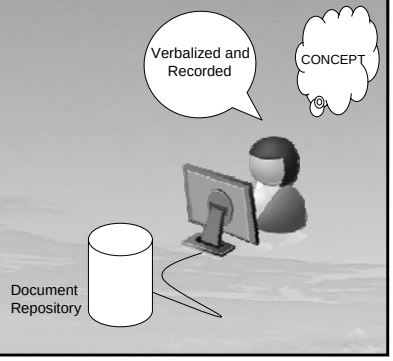
## Two Types of Resources

- Machine Utilization
  - Type and location of resource
  - Metadata about the resource
  - Low level semantics (for machine use)
- Metadata about the content in the resource
  - Describe content of document
  - Keywords and descriptive cataloging
  - High level semantics (for human use)

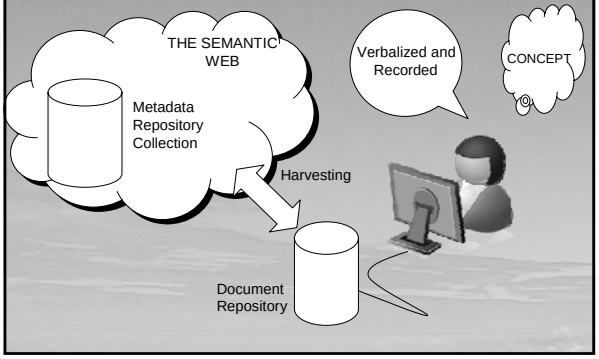
## Web-Based Information Life Cycle



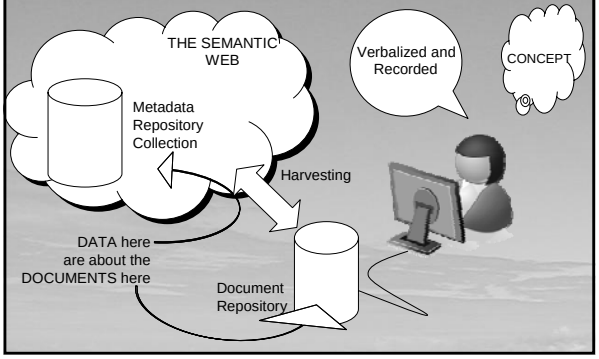

## Web-Based Information Life Cycle



## Web-Based Information Life Cycle




## Web-Based Information Life Cycle




## Metadata Repository

- Data about resource is explicit
  - Location
  - Type of resource
  - Other such data
- But it is limited
  - It is a representation
  - Representations do not capture all the information in the resource (keywords, etc.)
  - No information about scope of the representation: specificity and exhaustivity

## Temporary Solution: A Controlled Vocabulary

- A controlled vocabulary (CV) is a set of semantic structures of language used to represent the information that is found within the documents
- The most used types of CV:
  - Taxonomy: a classification scheme of the domain
  - Thesaurus: A semantically linked dictionary of terms that pertain to the domain
  - Ontology: Semantic relationships found in the domain
- The main problem is that not all concepts in the domain or in the documents get represented

## Multiple Repositories Compound the Problem

- Merging different data vocabularies is a challenge
- Merging different high-level semantic CVs
  - Assumptions about meanings
  - Assumptions about scope
  - Assumptions about audiences and users
  - Assumptions about uses, current and future
- Machine-side implementations work well only at the lowest level of lexical semantic

## Problems in Representation

- Controlled Vocabularies are useful to reduce the gap between high/human and low/machine levels of semantics but
  - Cannot capture all information
  - Cannot capture current information
  - Cannot foresee potential and future uses of the information
- They only offer a reduced view of the universe from their own perspective

## Languages and Initiatives

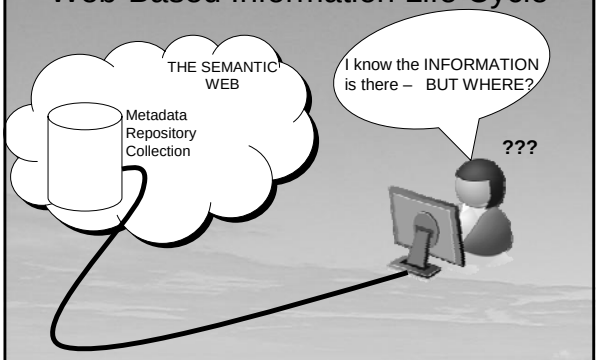
**LANGUAGES**

- XML family (XML, XLS, XPATH)
  - http://www.w3.org/XML/
- OWL
  - http://www.w3.org/TR/owl-guide/
- DAML + OIL
  - http://www.daml.org/language/

**INITIATIVES**

- Open Archives Initiative (OAI) – Protocol for Metadata Harvesting (PMH)
  - http://www.openarchives.org/
- Encoded Archival Description (EAD)
  - http://www.loc.gov/ead/
- Text Encoding
  - http://www.tei-c.org/
- MODS (Metadata Objects Description Schema) and METS (Metadata Encoding and Transmission Standards)
  - http://www.loc.gov/standards/mods/
  - http://www.loc.gov/standards/mets/
- Dublin Core metadata
  - http://dublincore.org/

## Web-Based Information Life Cycle




## Summary (1)

- Language is a communication tool but rich verbal interactions require a rich vocabulary
- M2M interactions requires strong formalization
- The standardized rules of the SW creates a rich environment for M2M interactions
- H2M and M2H interactions are not straightforward but require some formalization
- Matching the high-level semantics of human cognitive processes to the low-level of the semantics of the computer-based processes is not a trivial problem

## Summary (2)

- Applications to assist H2M and M2H interactions
  - Thinking, conceptualization and verbalization aids
    - Use of thesaurus and other semantic tools in word processing applications
  - Metadata
    - Creation and management
    - Merging and expanding schemes
  - Description of document content
    - Analyzers, language advisors
  - User-assisted document and information retrieval systems
    - Interactive feedback
    - Document filtering and summarization
    - Information analysis and discovery

## Summary (3)

- Observations
  - Overconfidence in computer-based systems
  - Lack of understanding of the limitations that systems have to treat high level knowledge (information = data in context)
  - Lack of end-user access to the system's infrastructure
  - Web Services and Controlled Vocabularies are not universal, they only exist within closed domains

## Summary (4)

- Creation of a CV is a high-level semantic task, it cannot be done as a purely automatic process, it requires human participation
- Mapping or merging multiple CVs require high-level semantic decision – it cannot be done as an automatic M2M process
- CVs are not static but dynamic and thus require monitoring, evaluation and maintenance – it cannot be done as an automatic M2M process either

## Summary (5)

- Web Services require formal semantic where data is unambiguous and rigid over time
  - Within restricted vocabulary, policies and implementations
- Some solid WS applications
  - Predictable business processes
  - Resource discovery
  - Directories
  - RDF
  - Database access and interactive operations

## Summary (6)

- Challenges
  - Multi-lingual and cross-lingual interactions
  - Cross-domain resource discovery
  - New functional demands
  - Trust
  - Stress on human operations across the enterprise
  - What does semantic interoperability mean?

## Case Study: OSCE

Organization of the Security and Cooperation in Europe

## Case Study (1): The Problem

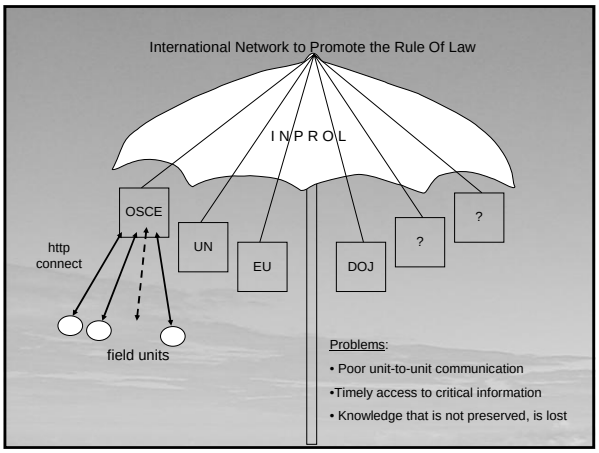
- **It exists in the RoL field operations**
  - There is no horizontal communication among field units
  - Timely access to critical information
  - Knowledge acquired in a field tour is not preserved due to missions being short term (6 months with the possibility of only a second 6-month extension)

## Case Study (2): A Solution

- **A centralized information system to serve several purposes**
  - A repository of best practices
  - Indirect means of communication among field units through sharing of experience
  - Organizational knowledge is transferred from one generation of officers to the next

## Case Study (3): Expansion

- **Similar projects are being addressed by other security organizations**:
  - UN – United Nations
  - EU – European Union
  - DOJ (homeland security)
  - Other yet unknown agencies




## Case Study (3): Current Challenge

- **The problem is not the technology but understanding how to manage information**
  - It is not interconnectivity of networks and of computer technology
  - It is the organization and the management of information resources (H2M, M2H, and H2H):
    - Users and user-to-computer interactions
    - Documents and collections
    - Knowledge
    - Information processes
    - Facilitation of knowledge retention

Thank you very much!

Questions?