行政院環境保護署
Environmental Protection Administration
Executive Yuan, R.O.C. (Taiwan)

An Ontology-Driven Approach for Harmonizing and Integrating Environmental Information

Yu-Chi Chu — Dept. of Environmental Monitoring and Information Management, **EPA, TAIWAN**

Su-Mei Huang, **Chin-Cheng Lien** — Dept. of Computer Science, SooChow University, **TAIWAN**

Chen-Chau Yang — College of Electrical and Computer Engineering, St. John University, **TAIWAN**

EnviroInfo 2007 Conference
Warsaw, Poland
September 12-14, 2007

Introduction

- Government agencies and private sectors need detailed information regarding the current state of the environment
- It is very difficult to share environmental data
  - resides on disparate databases
  - heterogeneity (**syntax** and semantics)
- There is a demand for appropriate systems to provide integrated environmental information

Heterogeneity issues

- **Structural heterogeneity**
  means that different information systems store their data in different structures.
- **Semantic heterogeneity**
  considers the **content of** an information item and its intended meaning.

Example -1

S_1
- Factory(facId, name, address, …)
- Permit(facID, permitNo, description, …)

S_2
- Plant(serialNo, plant-name, Plant-location, …)

S_3

```
<Site>
    <Name> … </Name>
    <Location>
        <latitude> …  </latitude>
        <longitude> …  </longitude>
    <Location>
    <WasteItem> … </WasteItem>
        .
        .
</Site>
```

- the offices in EPA do not all use the same database schema
- with the growing amount of data on the Internet, we are facing with data that is not well designed but with little structure such as HTML pages and XML documents

Example -2

There are a number of thesaurus, but without harmonization

Ontology

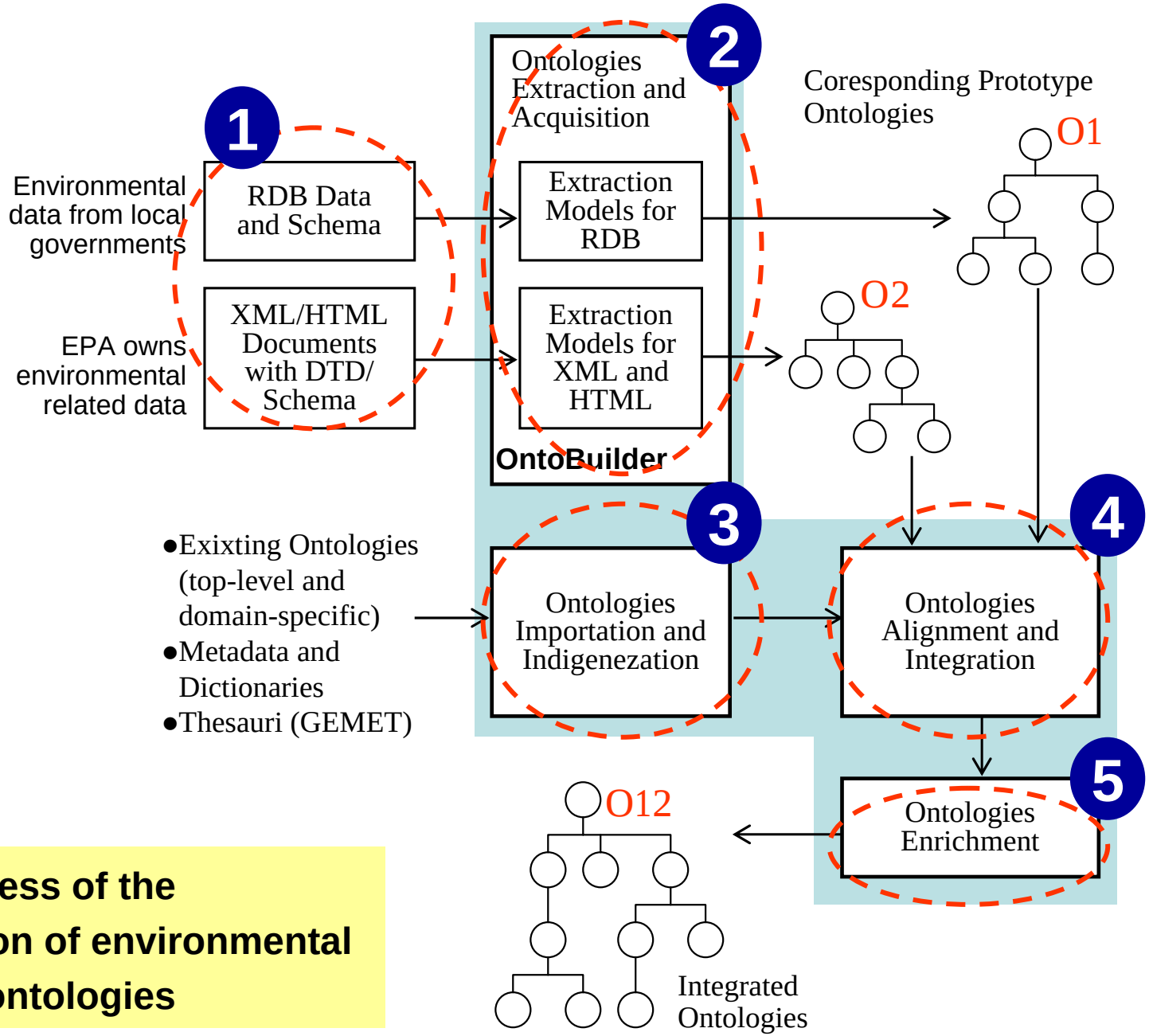
- Ontologies have gained importance in many areas of applications
  - interoperability, knowledge management
  - design of intelligent systems
- Used for a more precise specification of the semantic information content of the underlying data
- Domain specific ontologies are also being viewed as vehicles of capturing semantic information content

Methodology (1/2)

- Purpose and principles
  - ontologies to be practical and useful artefacts
  - the effort required to construct new ontologies must be minimized
  - overall effort required to construct an ontology must be amortized over multiple uses and users

Methodology (2/2)

- First, we provide an environment
  - assists users to build **prototype ontologies**
  - extracts knowledge from the existing information sources such as DB schemas or XML documents
- Secondly, for the built prototype ontologies
  - align with existing common ontologies and some of the domain specific thesauri such as GEMET
  - provide a semantic consistency among different ontologies
- Thirdly, the aligned ontologies will be merged and combined to form an **integrated ontology**



The process of the integration of environmental domain ontologies

Systematic processes (1/5)

- Selection of underlying information sources, standards, laws and regulations, classifications, etc.
  - select the sources of information relevant to the target domain
  - usually provide **taxonomy of concepts** and terminologies used in the domain

Systematic processes (2/5)

- Ontology extraction and acquisition
  - knowledge acquisition from the sources of information
  - form a prototype ontology for each knowledge source
  - employ the procedure of "**protocol analysis**" with domain experts
    - asking users to describe various types of domain applications, the data used in such applications, and the terms used in their field

Systematic processes (3/5)

- Ontology importation and indigenization
  - import ontologies that are existing ontologies in target domains but might usually be used in other countries or areas
  - tailor them to fit the feature in the area or country where ontologies will be applied
  - For example:
    - OECD classification
    - GEMET thesaurus

→ localized to fit in other countries or regions

Systematic processes (4/5)

- Ontology alignment and integration
  - align the prototype ontologies with imported ontologies or **upper level ontologies**
    - adjusting the name or terminology in the prototype ontologies
    - making them consistent with each other
  - combine and merge the prototype ontologies to form an integrated ontology

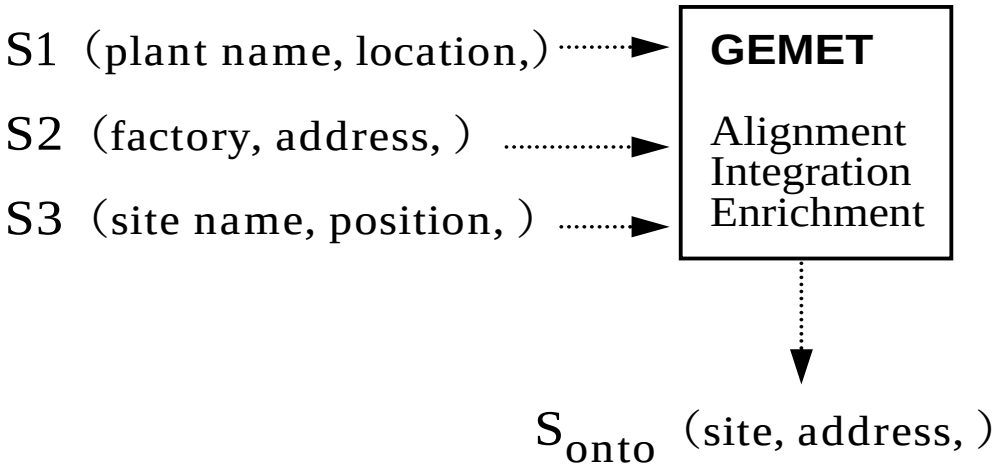
Systematic processes (5/5)

- Ontology enrichment
  - most of the ontologies merely represent taxonomy of concepts, where others may just include some attributes for them.
  - In this activity, we will enrich the integrated ontologies with extra information where available (attributes, features, …)

Example application

- Environmental Data Repository (EDR)
  - an integrated data warehouse system that provides a single point of access to data extracted from several major Taiwan EPA databases
  - construct the integrated ontology by extracting the domain knowledge from some of the information sources and aligning the concepts in the ontology with the laws and regulations of Taiwan EPA

an example of term integration



- each information source might maintain the data regarding the potential pollution sites
- conflict between the terms that is used to identify the sites
- ontology can assist us in overcoming this conflict, and develop a **consistent view** through which information can be integrated

EDR Homepage (http://edb.epa.gov.tw)

Conclusions

- a methodology based on ontology-driven related technologies to integrate environmental information
- It is shown, to some degree, that ontology can provide assistance in solving the heterogeneous problems among diverse information sources
- The approach may serve as an infrastructure component for integrating environmental data with known, but differing, collections of data

Future work

- recent advancements including:
  - Web services,
  - Ajax and Web 2.0 related technologies
  - Knowledge management

  might be integrated with the proposed approach
- design and implement a more sophisticated and practical system

Thank you!