ONTOLOGY-BASED MEDIATION OF OGC CATALOGUE SERVICE FOR THE WEB

A Virtual Solution for Integrating Coastal Web Atlases

Yassine Lassoued
Coastal and Marine Resources Centre, University College Cork, Naval Base - Haulbowline, Cobh - Co. Cork, Ireland
y.lassoued@ucc.ie

Dawn Wright
Department of Geosciences, Oregon State University, Corvallis, Oregon, USA
dawn@dusk.geo.orst.edu

Luis Bermudez
Southeastern Universities Research Association (SURA), Washington DC, USA
bermudez@sura.org

Omar Boucelma
Laboratoire des Sciences de l'Information et des Systèmes, Avenue Escadrille Normandie Niemen, 13397 Marseille, France
omar.boucelma@lsis.org

Keywords: Data Semantics, Semantic Web Technologies, Information retrieval, Geographic Information Systems (GIS), Ontologies, Catalogue Service for the Web (CSW), Mediation.

Abstract: In recent years significant momentum has occurred in the development of Internet resources for decision makers and scientists interested in the coast. Chief among these has been the development of coastal web atlases (CWAs). While multiple benefits are derived from these tailor-made atlases (e.g., speedy access to multiple sources of coastal data and information), the potential exists to derive added value from the integration of disparate CWAs, to optimize decision making at a variety of levels and across themes. This paper describes the development of a semantic mediator prototype to provide a common access point to coastal data, maps and information from distributed CWAs. The prototype showcases how ontologies and ontology mappings can be used to integrate different heterogeneous and autonomous atlases, using the Open Geospatial Consortium's Catalogue Services for the Web.

1 INTRODUCTION

The vast and heterogeneous amount of geospatial data on the World Wide Web causes users to be information overloaded (Kashyap and Sheth, 2000). Search engines return millions of results, including non relevant information, that are rarely taken into account. A user (e.g. scientist or a coastal response manager) would like to have as much relevant information as possible integrated for a particular event and region. This requires 1) providing a mechanism to allow unified discovery and access to distributed and heterogeneous data and 2) categorizing the results in a convenient vocabulary to the end user.

Practically, data discovery relies on documentation provided as part of metadata (notably discovery metadata) such as the dataset title, abstract, extent, keywords, etc. In the context of distributed resources, this information is present in different heterogeneous formats, and systems, according to several existing metadata models and standards.

The International Standardization Organization (ISO) has recently defined metadata standards for geospatial data, notably the ISO-19139 (ISO, 2006) standard. The aim is to harmonize metadata representation and implementation by conforming to a unified model, a unified structure and a unified format. The Open Geospatial Consortium (OGC), in their turn, focus on developing standards for querying and transporting data and metadata over the Internet. Specifically, the OGC Catalogue Service for the Web (CSW) (Nebert and Whiteside, 2005) specification defines a standard for advanced querying and transporting of metadata records over the Web.

Despite these harmonization efforts, problems still arise when dealing with metadata semantics. Terminology used to describe similar data can vary between specialities or regions, which can further complicate data searches and integration. For instance, usage of of the word "seabed" in Europe versus use of the word "seafloor" to describe the same feature in North America is a good example of this scenario, as is the interchangeable use of "coastline" versus "shoreline" in both regions. From both human and computational standpoints, users need assurance that the concepts, terminology, and even the abbreviations that are shared between two or more individuals, systems, or organizations are understood by all to mean the same thing. In this way the quality of data retrieval and subsequent data integration are greatly increased.

In this paper, we describe an ontology-based mediation approach for performing geospatial data search across different organizations. We use ontologies as the means to define semantics for metadata values (terms such as keywords, places, etc.) within organizations, but also to link terms from different organizations. An organization or a group of organizations populate their metadata using a local CSW. Their metadata records use a given ontology of terms called *local ontology*. Human or machine users formulate CSW requests using a common ontology of metadata terms, called *global ontology*. A CSW mediator rewrites the user's request into CSW requests over local CSWs using their own (local) ontologies, collects the results and sends them back to the user.

The paper is organized as follows. In section 2 we sketch the problem through a concrete integration example, while in section 3, we present state of the art interoperability techniques and technologies. In Section 4 we detail our CSW mediation approach, and in section 5 we present the implemented prototype. Finally, we conclude in section 6.

2 MOTIVATING EXAMPLE

The example described in this article is drawn from the coastal web atlases (CWAs) integration problem addressed by a new International Coastal Atlas Network (ICAN) initiative (Wright et al., 2007).

A CWA is a Web geographic information system composed primarily of coastal GIS data (vector data, coverages, raster grids, and images), their associated metadata, and thematic information about data (such as textual descriptions, references, images, etc.).

Integrating several CWAs requires three different levels of integration: integrating the GIS data, integrating their metadata, and integrating the thematic information that accompany data. The problem we are focusing on in this article is metadata interoperability, in other terms performing data discovery and search across different CWAs in a transparent way.

We report here on the development of a prototype as a proof-of-concept to inter-relate metadata between two initial CWAs: the Marine Irish Digital Atlas or MIDA, [http://mida.ucc.ie], and the Oregon Coastal Atlas or OCA, [http://www.coastalatlas.net]. It may not be immediately obvious how Oregon and Ireland may need to be interoperable, but these two mature atlas efforts can be used as a testbed for interoperability.

Both MIDA and OCA atlases implement the OGC CSW for querying and delivering metadata records. Metadata records in each of the CSWs use a given local ontology (e.g. keywords, places, titles, etc.). In order to facilitate search across both atlases, a centralized system needs to be implemented, which will provide unified and transparent access to the local CSWs. Ideally, metadata records from both MIDA and OCA will not be copied at the integrated level as both atlases are autonomous and are subject to evolution. Rather the integration system will act as a mediator (Wiederhold, 1992) that uses a common terminology (metadata ontology) and will translate user queries, on the fly, into queries over the atlases' CSWs using their own ontologies.

3 STATE OF THE ART

This section presents the state of the art of existing techniques and technologies that are related to the problem of semantic interoperability of GIS metadata management systems. The following subsections will explore the use of CSW to facilitate syntactic interoperability (c.f. subsection 3.1) and the use of mediation as integration approach (c.f. subsection 3.2).

3.1 OGC Catalogue Service for the Web (CSW)

CSW (Nebert and Whiteside, 2005) is an OGC abstract specification for supporting the ability to publish and search collections of descriptive information (metadata) for data, services, and related information objects. CSWs allow a unified access to metadata records within a community or an organization, thus harmonizing GIS resources discovery and search. For this reason, CSWs are required for coastal atlases in order to facilitate syntactic and schematic interoperability.

CSWs support several operations. We focus here on the `GetRecords` operation for searching metadata records, possibly using filters, such as keyword, location and time search.

There exist several implementations of CSW. For instance, both MIDA and OCA use GeoNetwork [http://geonetwork-opensource.org] as a CSW implementation.

3.2 Mediation

The database (DB) community has extensively studied and developed data integration approaches and systems leading to, among others, a virtual approach to data integration called mediation (Wiederhold, 1992).

A mediation system provides the user with a uniform interface of the different data sources via a common model. In a typical mediation architecture (c.f. Figure 1), several distributed data sources use their own data schemas, called *local schemas* or *source schemas*. Users pose queries over a common reference schema, called *global schema*. The mediator uses mapping rules between the global schema and the local schemas in order to rewrite the user's query into queries over the local data sources. It extracts and reformulates the responses conforming to the global schema and combines them in order to construct a response which is as complete as possible. Mediators often use a wrapper for each data source, for translating queries from the mediator's query language into the data source's query language, and for converting the source's data into the mediator's data model. All is transparent to the user. That is, the user ignores where and how data are stored and how the mediator manages to retrieve data from their sources.

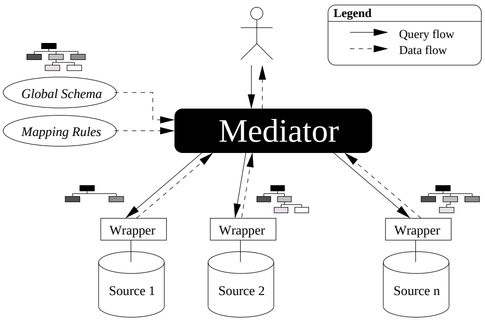

Figure 1: Typical Mediation Architecture

Several mediation systems and prototypes have been developed: examples of such systems are Ontomet (Bermudez, 2004), TSIMMIS (Garcia-Molina et al., 1997), PICSEL (Goasdoue et al., 2000), Information Manifold (Kirk et al., 1995), AGORA (Manolescu et al., 2001). Most of these existing mediation approaches focus on data schemas heterogeneity and data complementarity and try to build responses which are as complete as possible.

4 CSW ONTOLOGY-BASED MEDIATION

The targeted integrated Coastal Web Atlas, called *global*[1] *atlas* or *super atlas*, is a virtual atlas that offers transparent access to a variety of distributed and heterogeneous local coastal atlases. The notion of "virtual", in this context, means that local atlas resources are not integrated or copied at the integrated level. Rather, they remain at their locations and are remotely accessed, harmonized and integrated on the fly depending on users' requests. This allows a high degree of independence and autonomy for the local atlases and facilitates extendibility in an architecture where atlases can be added and removed at any time without affecting the global atlas, provided that they implement core services including OGC CSW for the delivery of metadata, Web Map Services (WMS) (de la Beaujardiere, 2004) for the delivery of maps, and WFS (Vretanos, 2005) for the delivery of vector data.

In this article, we focus on the data discovery and search aspects. We propose an ontology-based mediation approach for OGC CSWs. The solution differs from the classical mediation approaches cited in subsection 3.2 in the way that it deals with metadata which are already in the same format, XML, and have the same ISO-19139 schema. It does not try to combine information from different sources (i.e. atlases or CSWs); rather it focuses on the semantic values contained within the metadata records and tries to solve semantic conflicts between different applications, domains, organizations, or simply CSWs. Our approach is ontology-based, i.e. it uses ontologies for representing the semantics of data values as well as for matching concepts of local ontologies with concepts of the global ontology.

4.1 Architecture

The global atlas introduced above offers a virtual CSW, called global CSW, which acts as a CSW mediator and which offers unified and transparent access

[1] Please note that the term "global" does not refer to the globe in this context. Rather, it is the term used by the database community to refer to the integrated data schema in a mediated approach as opposed to local schemas.

to the atlases' CSWs. As illustrated in Figure 2, users of the global CSW are provided with a global ontology of terms[2]. The user refers to the global ontology and formulates a CSW `GetRecords` request (c.f. subsection 3.1) using an area of interest and keywords defined in the global ontology. The global CSW rewrites the user's request into CSW requests over the local atlases' CSWs using their local ontology terms, executes the so-obtained requests, and collects metadata records (responses) from local CWAs.

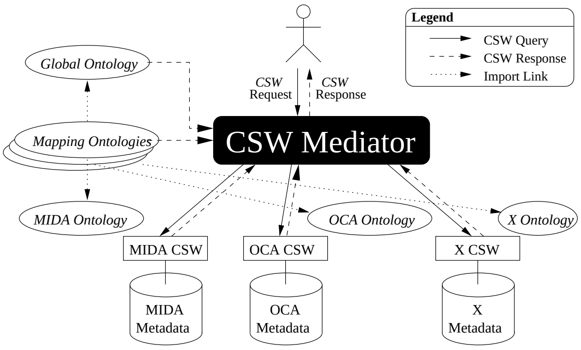

Figure 2: Ontology-Based CSW Mediation Architecture

This architecture facilitates extensibility as new catalogue services can be added and removed at any time without affecting the global CSW, provided that they come with the ontologies for the terms used by their metadata records and that mappings between these terms and the global ontology's terms are provided. Another advantage of this architecture is that the global CSW acts itself as a catalogue service, which in its turn can be queried by another external application or even integrated in a similar CSW mediation architecture, as a local CSW.

4.2 Global and Local Ontologies

A (global or local) CSW uses an ontology which defines the terms used as values in its metadata records (for example thematic keywords, places, etc.). In the initial CSW mediator prototype, we only focus on values for keywords provided as part of metadata. Conforming to the ISO-19115 standard, five types of keywords are defined: *discipline*, *theme*, *place*, *temporal*, and *stratum*. For each atlas, an ontology of terms related to these five keyword types is defined. Relationships between the terms contained within one ontology are provided as part of the same ontology.

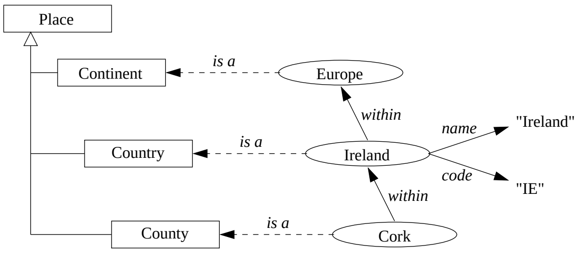

Figure 3: Place terms from the MIDA ontology

Figure 3 illustrates an extract of place keywords from the MIDA ontology. Examples of places are respectively Europe, Ireland and Cork. Relationships between places such as "Cork *is within* Ireland", and "Ireland *is within* Europe", can be expressed. This helps improve keyword search using an inference engine. For example, if the place keywords for a dataset only contain the term "Cork", and a user queries the metadata catalogue using the place term "Europe" or "Ireland" they still will get this dataset in the response, because Cork is in Ireland and Ireland is in Europe, which also infers that Cork is in Europe.

Ontologies are expressed in the OWL-DL language (Herman, 2007) in the ICAN CSW mediator prototype.

4.3 Ontology Mappings

Ontology mappings link the global ontology to the local ontologies. This link is crucial as it is the only means to allow the CSW mediator rewrite user requests expressed with terms from the global ontology into requests over the local CSWs using their own terms. Therefore, they act, as semantic translators.

For each local ontology, an OWL ontology called mapping ontology defines the mappings between the local ontology and the global one. A mapping ontology imports both a local and the global ontology and defines relationships between their concepts. An example showing extracts of the MIDA and the OCA mapping ontologies is illustrated in Figure 4.

In Figure 4, terms preceded by prefix "*global*" are from the global ontology. Those preceded by prefixes "*mida*" and "*oca*" are respectively from the MIDA and OCA ontologies. Relationships represented with thin lines are defined as part of the local ontologies. Those represented with thick lines are defined as part of the mapping ontologies. For example, *Coastal Protection* and *Shore Stabilization* are defined in the MIDA and OCA mapping ontologies as narrower terms than *Human Responses to Coastal Change*.

[2] The current use case topic that the ICAN group is focussing on is coastal erosion

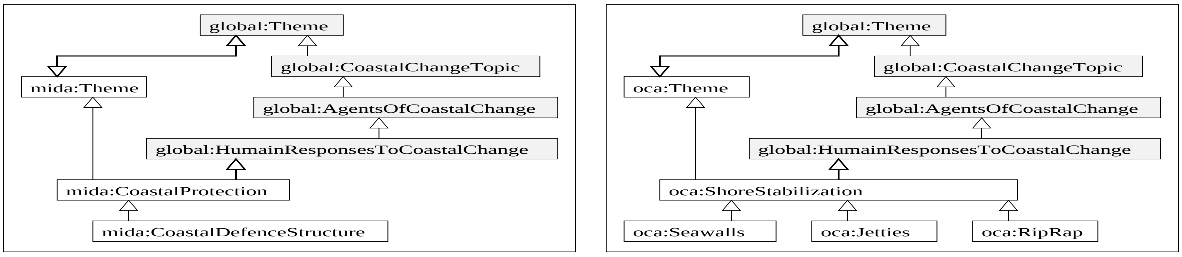
global:Theme
global:CoastalChangeTopic
mida:Theme
global:AgentsOfCoastalChange
global:HumainResponsesToCoastalChange
mida:CoastalProtection
mida:CoastalDefenceStructure
global:Theme
global:CoastalChangeTopic
oca:Theme
global:AgentsOfCoastalChange
global:HumainResponsesToCoastalChange
oca:ShoreStabilization
oca:Seawalls
oca:Jetties
oca:RipRap

Figure 4: Extracts of the MIDA and the OCA Mapping Ontologies

4.4 Query Rewriting and Execution

Query rewriting is the most important task of the mediator as it refers to the process of rewriting a user query posed over a global schema into queries over local data sources (in this context CSWs). We mainly focus on the GetRecords CSW requests in this section. Consider a human or software user that poses a GetRecords request, searching for metadata records available through the global CSW. For instance, the following is a GetRecords request formulated by a user interested in metadata records about data covering any region all over the world and related to *Human Responses to Coastal Change*.

```
http://ican.ucc.ie/srv/en/csw?
request=GetRecords&service=CSW&version=2.0.1
...
&constraint=
<?xml version="1.0" encoding="UTF-8"?>
<Filter xmlns="http://www.opengis.net/ogc"
    xmlns:gml="http://www.opengis.net/gml"
    xmlns:csw="http://www.opengis.net/cat/csw/2.0.2">
  <And>
    <PropertyIsLike wildCard="%"
     singleChar="_" escape="\">
      <PropertyName>keyword</PropertyName>
      <Literal>
         HumanResponsesToCoastalChange%
      </Literal>
    </PropertyIsLike>
    <BBOX>
      <PropertyName>
        /csw:Record/ows:BoundingBox
      </PropertyName>
      <gml:Envelope
       srsName=
       "http://www.opengis.net/gml/srs/epsg.xml#4326">
        <gml:lowerCorner>-180 -90</gml:lowerCorner>
        <gml:upperCorner>180 90</gml:upperCorner>
      </gml:Envelope>
    </BBOX>
  </And>
</Filter>
```

As we only deal with keywords semantics in this paper, the query rewriting process is quite simple. The process consists in translating the global keywords contained in the user's queries into local keywords and rewriting the initial request using the so-obtained terms. In order to do so, the mediator starts by parsing the user's request. It identifies the clauses related to keywords, and extracts the corresponding keyword literals. For instance, in the example above, the only clause related to keywords is the one delimited by the first <PropertyIsLike> tag, and the corresponding keyword is *Human Responses to Coastal Change*.

For each local atlas, the CSW mediator uses its inference engine to obtain all the local atlas' terms that are equivalent to, or narrower than, the keyword literal considered. Next, the initial clause is replaced by a disjunction of clauses, each containing a keyword literal corresponding to one of the so-obtained local keywords. For instance, in the example above, the CSW mediator will translate keyword *Human Responses to Coastal Change* into the MIDA keywords *Coastal Protection* and *Coastal Defence Structure*. Thus the corresponding clause will be rewritten according to MIDA as follows:

```
<Or>
  <PropertyIsLike wildCard="%" singleChar="_"
   escape="\">
    <PropertyName>keyword</PropertyName>
    <Literal>
      CoastalProtection%
    </Literal>
  </PropertyIsLike>
  <PropertyIsLike wildCard="%" singleChar="_"
   escape="\">
    <PropertyName>keyword</PropertyName>
    <Literal>
      CoastalDefenceStructure%
    </Literal>
  </PropertyIsLike>
</Or>
```

This process is repeated for each clause in the request's filter, of course by avoiding repetition of keywords in a disjunction of clauses. Each so-obtained final GetRecords request is sent to the corresponding local CSW. Records obtained as results from the local CSWs are then collected and sent back to the user.

5 IMPLEMENTATION

A first version of the global coastal atlas prototype has been implemented in Java, using the Jena 2 framework for inference purposes and is available at [http://ican.ucc.ie]. The prototype allows the user to:

- Select keywords from the global ontology;
- Select an area of interest;
- Submit a query, which will generate a CSW `GetRecords` requests to the global CSW.

The CSW mediator will consult a registry of atlases and identify the atlases that may have data within the bounding box selected, as a bounding box is associated with each atlas representing its geographic extent in order to optimize query execution by avoiding rewriting queries over CSWs with no data covering the area of interest. Next, the atlas mediator will translate the user's request according to the process described in subsection 4.4, collect the responses and send them back to the user through the graphical interface.

6 CONCLUSIONS AND FUTURE WORK

The atlas mediator prototype described in this paper is a first step towards atlas integration as part of a new International Coastal Atlas Network (ICAN). The prototype showcases how ontologies and ontology mappings can be used to integrate different heterogeneous and autonomous atlases (or information systems), particularly OGC CSWs.

The next step of the ICAN initiative is to integrate WFS and CSW mediation techniques in order to define a more complete approach for integrating both data and metadata. Also, thematic information will be considered and interfaces will be specified for sharing this type of information, which is highly important in atlases. The aim is to define a complete solution for integrating CWAs.

An initial evaluation revealed that the number of inferred keywords in rewritten queries can be dramatic depending on how general the user's selected keywords are. In some cases, one global keyword can correspond to more than sixty local keywords of a local atlas. This results in huge queries whose execution can be time consuming, especially when a large number of users are connected to the global atlas at the same time. In addition, no effort has been made to rank metadata records according to relevance, date, or spatial proximity. Future work will take these problems into consideration in order to optimize query rewriting and execution as well as results presentation.

REFERENCES

Bermudez, L. E. (2004). *Ontomet: Ontology Metadata Framework*. PhD thesis, Drexel University, Philadelphia, USA.

de la Beaujardiere, J. (2004). *OGC Web Map Service Interface (Version 1.3.0)*. Open Geospatial Consortium Inc.

Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaman, A., Sagir, Y., Ullman, J., Vassalos, V., and Widom, J. (1997). The TSIMMIS approach to mediation: Data models and Languages. *Journal of Intelligent Information Systems*.

Goasdoue, F., Lattes, V., and Rousset, M.-C. (2000). The Use of CARIN Language and Algorithms for Information Integration: The PICSEL System. *International Journal of Cooperative Information Systems*, 9(4):383–401.

Herman, I. (2007). *Semantic Web – Web Ontology Language (OWL)*. W3C.

ISO (2006). *ISO/PRF TS 19139 – Geographic information – Metadata – XML schema implementation*. International Standardization Organization.

Kashyap, V. and Sheth, A. (2000). *Information Brokering Across Heterogeneous Digital Data – A Metadata-based Approach*. Kluwer Academic Publishers, Norwell, Massachusetts, USA.

Kirk, T., Levy, A. Y., Sagiv, Y., and Srivastava, D. (1995). The Information Manifold. In Knoblock, C. and Levy, A., editors, *Information Gathering from Heterogeneous, Distributed Environments*, Stanford University, Stanford, California.

Manolescu, I., Florescu, D., and Kossmann, D. (2001). Answering XML Queries over Heterogeneous Data Sources. In *Proceedings of VLDB*, pages 241–250.

Nebert, D. and Whiteside, A. (2005). *OGC Catalogue Services Specification*. Open Geospatial Consortium Inc.

Vretanos, P. A. (2005). *Web Feature Service Implementation Specification*. Open Geospatial Consortium Inc.

Wiederhold, G. (1992). Mediators in the Architecture of Future Information Systems. *IEEE Computer*, pages 38–49.

Wright, D., Watson, S., Bermudez, L., Cummins, V., Dwyer, N., O'Dea, L., Nyerges, T., Benoit, G., Berman, M., Helly, J., and Uhel, R. (2007). Report on Coastal Mapping and Informatics Trans-Atlantic Workshop 2: Coastal Atlas Interoperability. Internal, unpublished workshop proceedings.