Ontology based Automatic ETL for Marine Geoscientific Data

T.Naz[1], Y.Lassoued[2] and R.Wallace[1]

[1]*Cork Constraint Computational Centre, University College Cork, Cork, Ireland*
Email: t.naz@4c.ucc.ie
[2]*Coastal and Marine Resources Centre, University College Cork, Cork, Ireland*

Abstract

Large volumes of marine geoscientific datasets have been gathered by various institutes over the past number of years. In order to add value to these very costly and valuable products and "improve the quality of scientific advice," an effort must be spent on providing integrated management and access to these datasets. This will allow a more holistic or "ecosystem" approach in the analysis of marine and geoscientific data. The objective of the GeoDI (Geological & Geophysical Data Integration) project is to derive maximum value from the national data acquisition effort to date and to allow future data to be integrated easily. As part of GeoDI a database is being designed and implemented for integrating marine geoscientific datasets using a common structure and common semantics. A key issue that is addressed by GeoDI is populating the database using the datasets that are continuously being collected. As a matter of fact, data collection procedures are continuously evolving, resulting in a variety of data formats, structures and semantics. GeoDI is designing and developing an automatic ontology-based ETL system for marine geoscientific data. The system automatically (i) extracts the structure and semantics of a new dataset to be integrated, (ii) matches the dataset structure and semantics to those of the integrated database, (iii) transforms the dataset according to the integrated database schema, and (iv) loads it. The GeoDI ETL system uses ontologies as a way to represent data structure and semantics. It is based on an extensible multi-strategy learning approach wherein different matchers (learners) are trained separately to match new schemas to the integrated database schema. Given a new dataset to be integrated into the geoscientific database, each learner maps the schema of the dataset to that of the integrated database. Decisions of the various learners are then combined by a meta-matcher.

Keywords: Marine Geoscientific data, Ontology based integration, ETL, GeoDI ETL, Data integration system

1 Introduction

Large volumes of marine geoscientific (geological/geophysical) datasets are gathered by different institutes during different surveys using a variety of instruments, and are stored in a variety of formats and representations. The collected datasets in the Arc marine domain are heterogeneous in nature, and there is considerable heterogeneity including differences in format, syntactic features and semantic interpretations across

geo-scientific datasets. Different geo-scientific datasets may be stored in CSV, Shape, MS Access, MS Excel, XML, MySQL etc formats thus showing format heterogeneity. Different schemas show syntactic and semantic heterogeneity (see Example 1 and 2) and there is a lack of common structure and common semantics to represent/integrate the marine geoscientific datasets. For GeoDI project, there also exist map cardinality problems i.e. 1:1 and 1:n (see Example 3 and 4).

Example 1: In the Arc Marine data model for geoscientific datasets, one survey schema may use "ORGANICS_F" and another may use "Organics/Fossils" to represent the same category of information "Organic_Fossils".
Example 2: In the Arc Marine data model for geoscientific datasets, one survey schema may use "Vessel_COD" and other may use "Veh_ID" to represent the VehicleID.
Example 3: In the Arc Marine data model for geoscientific datasets, one survey schema may use "sample_ref_nu' and another may use "SRefNo to represent the same category of information. This is an example of a 1:1 mapping. Other dataset may use "Ref" or "Reference_No" etc to represent the same category of information.
Example 4: One schema may use "Name" and another combination of "First Name" and "Last Name" to represent the name. This is an example of a 1:n mapping.

Datasets in the marine domain show that there are different concepts and granularities of knowledge and schema/data integration in the marine domain is a challenging task. In this paper we describe GeoDI ETL (Extract, Transform and Load tool for Geological and Geophysical Data Integration) – a tool that can automatically extract marine geoscientific data from different formats and develop a mechanism to translate between different concepts from multiple schemas. We have developed a domain ontology in the marine domain as a way to represent data structure and semantics. The ontology and multi-strategy matchers (learners) are developed to translate the concepts related to datasets from the surveys, according to the integrated format. The translated concepts are helpful to store data in the central repository of GeoDI.

The rest of the paper is organized as follows. Section 2 reviews related work in the schema and ontology matching/mapping from the history that are pertinent to GeoDI project and existing ETL tools. Section 3 presents the GeoDI ontology. Section 4 presents our overall design for GeoDI ETL with the main components of a GeoDI ETL tool and shows how our tool works. Section 5 describes a short case study in geoscientific marine data. Section 6 discusses our contributions, and gives directions for future work.

2 Related Work

Researchers from schema/data matching and mapping area have made considerable efforts (Rahm and Bernstein, 2001; Madhavan *et al.*, 2001; Shvaiko and Euzenat, 2005; Hakimpour and Geppert, 2001; Embley *et al.*, 2004; Aumueller *et al.*, 2005; Karasneh *et al.*, 2009). Rahm and Bernstein (2001) investigates many prototype implementations and presents a taxonomy for existing schema matching approaches i.e. schema level and instance level, element level and structure granularity (including top down and bottom up approach), linguistic- based and constraint-based. Madhavan *et al.* (2001) proposes an algorithm Cupid that uses different approaches and utilizes name, data types, constraints, schema structure, linguistic matching, structural

matching, context dependent matching and leaf structure for schema matching. Shvaiko and Euzenat (2005) classifies and distinguishes between syntactic, semantic and external techniques at element and structure levels. Hakimpour and Geppert (2001) present an approach to integrate different schemas from different communities into a single global schema for federated database systems. Embley *et al.* (2004) introduces two matchers i.e. object-set and structure matchers to improve the matching process. COMA++ (COmbining MAtch) is a schema and ontology matching tool, supporting 15 matchers to identify semantic correspondences between meta-data structures or models (Aumueller *et al.,* 2005). Karasneh *et al.* (2009) utilizes five matchers i.e. relation schema matcher, attribute name matcher, data-type matcher, constraint matcher and instance data matcher to solve the problem of schema matching for heterogeneous relation databases. The authors claim that the process of schema matching is fully automatic without any human intervention and their approach has achieved higher percentage of similarities and percentage of matched attributes compared to the other approaches.

Research on a very large scale is in progress by the ontology community in the field of ontology matching and mapping (Euzaenat and Shavaiko, 2007). Wache *et al.* (2001) has analyzed 25 approaches using ontologies as a solution to semantic heterogeneity problem and information integration. Single ontology approach, multiple ontology approach or hybrid ontology approach is used for the identification and association of semantically corresponding information concepts. Naz and Dorn, (2009) proposed a hybrid approach for schema and data integration for meta-search engines and the integration is based on single domain ontology.

There exist many hand coded and tool based ETL tools. Open source ETL tools including Apatar, CloverETL, JitterBit 2.0, Pentaho Data Integration, Scriptella, Talend Open Studio, KETL, Jasper and a commercial tool Microsoft SQL Server Integration Services (SSIS), have been studied (Vassiliadis *et al.,* 2003; ETL, 2010; SQL-Server, 2010; Vivantech, 2010; GeoDI-UCC-D2.5, 2010). From GeoDI perspective, there exist some problems with license cost free ETL tools, code-generator ETL tools and engine-based ETL tools. Today's free ETL tools are quite suitable when they are used within limits and they are missing a) advanced connectivity, b) techniques to handle domain's complex transformations and c) techniques for complex data integration. The problem with code-generator ETL tools is that many transformations require manual coding and require in-depth knowledge of programming language. Some engine-based ETL tools require complex engine's configuration.

Based on the review of existing schema/ontology matching techniques and ETL tools, we decided to develop an ontology based automatic ETL tool for marine geoscientific data that use multi-strategy learning approach (multiple matchers) for schema/data integration.

3 GeoDI Ontology

GeoDI database is being designed and implemented for integrating marine geoscientific datasets using a common structure and common semantics. GeoDI ontology has been designed in OWL (Web Ontology Language) by using above

developed common structure and semantics. Ontology also contains some grouping information that helps the GeoDI ETL in schema transformation.

4 GeoDI ETL Design

Our GeoDI ETL process involves the three usual components namely the GeoDI Extractor, GeoDI Transformer and GeoDI Loader. Figure 1 shows the GeoDI ETL process. The GeoDI ETL process is based on semantic Web technologies. By "based on semantic Web technology," we mean that a domain ontology is used in the ETL process.

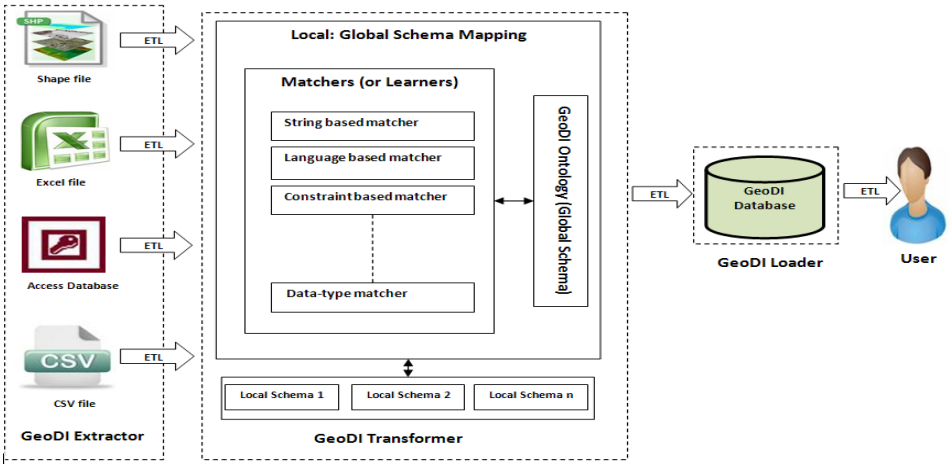

Figure 1. Ontology based ETL Process.

The GeoDI Extractor is responsible for extracting the schema from different data sources. The schema extracted by the GeoDI extractor is called the local schema. Then every local schema is given to the GeoDI Transformer. The GeoDI Transformer component is responsible for finding the automatic mappings between every local and global schema. In this case, the global schema, also known as model schema, is represented by the GeoDI ontology. GeoDI ontology is used as a way to represent data structure and semantics. During schema mapping different types of matchers (learners) are trained e.g. string-based matcher, language-based matcher and constraint-based matcher. For a schema matching process (transformation process), we use GeoDI domain ontology and trained learners to help us to find complex mappings. After the schema is successfully mapped to the global schema, the data is loaded to the GeoDI database with the help of GeoDI Loader.

The components of GeoDI ETL process are given below in more detail.

4.1 GeoDI ETL Extractor

The GeoDI-Extractor is responsible for extracting schema/data from different data sources i.e. ESRI shape files, MS Excel files, flat files and MS Access database. Format heterogeneity makes it difficult to integrate data so our GeoDI-Extractor component resolves the problem of format heterogeneity by accessing four data formats by using different APIs e.g. GeoTools for ERSI shape files (GeoTools, 2010),

Apache POI for MS Excel data sources (Apache POI, 2010) and java packages to extract the schema/data from the flat files and MS Access databases.

4.2 GeoDI ETL Transformer

As mentioned before, there exists syntactic and semantic heterogeneity between geo-scientific datasets. The GeoDI-Transformer component is responsible for resolving the syntactic and semantic heterogeneity. With the help of GeoDI Transformer, our key requirement is to provide automatic techniques for schema/data matching and integration, utilizing techniques derived from the database and ontology communities. Different data-sets/schemas use different names and there can be different data types as well. So automatic mapping discovery between local schemas from marine geoscientific domain and GeoDI model schema is not an easy task. As a result we need extensible multi-strategy learning approach for schema transformation process that can solve this complex problem. These matchers (learners) are trained separately to match local schemas to the integrated database. Decisions of various matchers are then combined by a meta-matcher.

4.2.1 Schema Level Matcher

GeoDI matchers are schema-based and use schema information i.e. names, data type, relationships and constraints etc to find a match between schemas. In the case of GeoDI, it is not possible to use instance-level technique that use data instances, since many data instances are numbers only and are not in a particular format or pattern. An instance-based matcher basically focuses on analysing the data values of attributes.

4.2.2 Element Level Matcher

The GeoDI element-level matcher computes a mapping between individual schema elements (pair of attributes), e.g., an attribute matcher by using string-based matcher and linguistics-based matcher.

Our string-based matcher uses a stemming algorithm and different string distance functions to find a similarity between strings. In particular, the porter stemming algorithm removes the prefix and suffix of a string, handles singular and plural of concepts, and then finds the similarity between strings (Porter, 2006). The following are example resolved with the porter stemming algorithm (Porter, 2006):

Organics/Fossils→Organic_Fossils, ORGANICS_ F→Organic_Fossils

We utilize two different string distance algorithms Jaro and Levenshtein distance (Chapman, 2006) for GeoDI transformer. If the sum of their similarity scores exceeds a threshold value, we consider this as a positive match. The following types of problems are resolved by using string distance functions.

SUB_SAMPLI→SubSampling, MUNSELL_CO→ MunsellColourCode,
Strength/Compactness (clay/slit) →Strength_Compactness,
Fabric/Microfabric→Fabric_Microfabric, Biogenic
content/shells→BiogenicContent

Our linguistics-based matcher is based on natural language processing techniques, including tokenization and elimination. Tokenization involves the removal of punctuation, blank spaces, and adjustment of cases. Elimination involves the removal of stop words (a list of stop words for the given domain needs to be provided to the system). In GeoDI case, stop words include cm, µm and % etc. The following type of problems are solved by string transformation using tokenization and elimination:

DATE_→Date, Ripple height (cm) →RippleHeight,
Mean (µm) →Mean_micrometer, Clay (%)→Clay_percent

4.2.3 Constraint-Based Matcher

The GeoDI constraint-based matcher use schema constraints, such as data types and intra-schema relationships such as referential integrity. Our constraint based matcher consists of data type matcher and relation schema matcher. The GeoDI data type matcher use a synonym table specifying the degree of compatibility between a set of predefined generic data types (Karasneh *et al.,* 2009).

The relational schema matcher compares two relational schemas Si and Sj of two different databases Di and Dj to identify similarities between these schemas. In GeoDI domain, if the schema is extracted from the MS Access relational database then we can utilize the relation schema matcher. In this case Si is table from sample MS Access database (Di) and Sj is a table from the GeoDI model (Dj) represented by a "Class" in ontology. The similarity between the names of tables is calculated by using string based matcher or linguistic based matcher described above (Karasneh *et al.*, 2009).

4.2.4 Match Cardinality

In GeoDI schema matching, we can discover 1:1 and 1:n mappings. The following type of 1:n mapping solutions are detected.

SIZE→ "MinSize" and "MaxSize",
AMOUNT→ "MinAmount" and "MaxAmount"

4.2.5 Combinational Matcher

GeoDI use multiple-strategy learning approach for the ETL tool. For GeoDI, we use a single ontology approach, and the GeoDI domain ontology act as a global ontology that represents the data structure and semantics. In a combinational matcher, the local schemas are matched to the GeoDI global schema (i.e. the GeoDI ontology) by using multiple matchers. It can use synonyms associated with concepts in the domain ontology, multiple similarity measure algorithms or any matcher defined above etc. The following are examples resolved by using a combinational matcher.

Water_Dept→WaterDepth, Vessel→Vehicle, Initial→Name Title,
Instrument→ Device or MeasuringDevice, Testing_code→TechniqueID

4.3 GeoDI ETL Loader

When the matching process for all attributes is completed by the GeoDI transformer the data is ready to send to the central repository. The GeoDI-Loader component is used to transform the data into GeoDI data warehouse. It is possible that we find multiple mappings for local schema elements by the GeoDI transformer or do not discover any mapping for the complex cases, in such situation user verification is required. If GeoDI Transformer proposes multiple mappings then user can choose any one (appropriate) from the list of suggested mappings and if there do not exist any mapping then user can select the mapping on his own from the ontology.

5 Case Study

S_1 is the schema collected from the Marine Institute (Galway, Ireland) and O_{GeoDI} is a model schema for marine geoscientific domain. S_1 contains worksheet named "Sediment Type" that contains the attributes {Ref No:, Sample Name, Date / Time, Mean (µm), Sorting, Clay (%), Silt (%), Mud (%) [cl+silt], Sand (%), Gravel (%), Check, Description, Comment on Upper Fraction}. The GeoDI ETL tool extracts the schema from MS Excel worsheet, suggests that data must be mapped to the "SedimentologicalAnalysis" entity of GeoDI data model. It also discovers the mappings between the schema S_1 and O_{GeoDI} model as below.

S_1. Mean (µm)→O_{GeoDI}.Mean_micrometer, S_1. Sorting →O_{GeoDI}. Sorting,

S_1. Clay (%) →O_{GeoDI}. Clay_percent, S_1. Silt (%)→O_{GeoDI}. Silt_percent ,

S_1. Mud (%) [cl+silt] →O_{GeoDI}. Mud_percent, S_1. Sand (%)→O_{GeoDI}. Sand_percent,

S_1. Gravel (%)→O_{GeoDI}. Graval_percent, S_1. Check→O_{GeoDI}. Check_percent,

S_1. Comment on Upper Fraction →O_{GeoDI}. CommentOnUpperFraction

6 Contributions and Future Work

In this paper we have proposed a GeoDI ETL that resolves the problem of schema/data matching and integration for marine geoscientific data. We have introduced a multi-strategy approach for the marine geoscientific domain that uses multiple learners to automatically populate the database by the datasets that are continuously being collected in various formats. We have also designed an ontology and database for integrating marine geoscientific datasets using a common structure and common datasets. In the future, we will introduce more type of file formats e.g. XML, SQL or MySQL to the extractor component. More learners can be introduced to improve the transformation process.

7 Acknowledgements

The work is supported by Marine Institute, Galway, Ireland under project PBA/KI/07/001.

References

Aumueller, D., Do, H., Massmann, S., and Rahm, E., 2005. Schema and ontology Matching with COMA++, *Proceedings of the 2005 ACM SIGMOD International Conference on Management Data*, pp. 906-908, Maryland, USA

Apache POI, 2010. Apache POI - the Java API for Microsoft Documents http://poi.apache.org/

Chapman, S., 2006. SimMetrics, http://www.dcs.shef.ac.uk/~sam/stringmetrics.html

ETL, 2010. ETL Enterprise Data Integration, http://www.etltools.net/

Euzenat, J., and Shvaiko, P., 2007. Ontology Matching, *Springer*, ISBN 978-3-540-49611-3

Embley, D. W., Xu, L., and Ding, Y., 2004. Automatic Direct and Indirect Schema Mapping: Experiences and Lessons Learned, *ACM SIGMOD Record*, pp.14-19

GeoTools, 2010. GeoTools The Open Source Java GIS ToolKit, http://www.geotools.org/

GeoDI-UCC-D2.5, 2010. Review of ETL and Ontology/Schema Matching Techniques and Tools.

Hakimpour, F., and Geppert, A., 2002. Global Schema Generation Using Formal Ontologies, *Proceedings of the 21st International Conference on Conceptual Modeling*, pp. 307-321

Karasneh, Y., Ibrahim, H., Othman, M., and Yaakob, R., 2009. Matching Schemas of Heterogeneous Relational Databases, *Proceedings of Second International Conference on the Applications of Digital Information and Web Technologies*, London, UK , IEEE Explore, pp. 1-7, ISBN: 978-1-4244-4456-4

Madhavan, J., Bernstein, P. A., and Rahm, E., 2001. Generic Schema Matching with Cupid, *Proceedings of the 27th VLDB Conference*, pp. 49-59, Roma, Italy

Naz, T., and Dorn, J., 2009. Configurable Meta-Search in the Human Resource Domain- A Hybrid Approach to Schema and Data Integration for Meta-search Engines, *Lambert Academic Publishing*, ISBN: 978-3-8383-0230-0

Porter, M., 2006. The Porter Stemming Algorithm, http://tartarus.org/~martin/PorterStemmer/

Rahm, E., and Bernstein, P. A., 2001. A Survey of approaches to Automatic Schema Matching, *VLDB Journal - The International Journal on Very Large Data Bases*, **10**(4): 334-350, ISSN: 1066-8888

Shvaiko, P., and Euzenat, J., 2005. A Survey of Schema-based Matching Approaches, *Technical Report DIT-04-087, Informatica e Telecomunicazioni*, University of Trento, Italy

SQL-Server, 2008. Microsoft SQL Server 2008, http://www.microsoft.com/sqlserver/2008/en/us/spatial-data.aspx

Vassiliadis, P., Simitsis, A., Georgantas, P. and Terrovitis, M., 2003. A Framework for the Design of ETL Scenarios, Advanced Information Systems Engineering book, *Springer*, ISBN: 978-3-540-40442-2

Vivantech, 2007. The Evolution of ETL - From Hand-coded to Tool-based ETL, http://vivantech.net/Documents/ETL-Evaluation.pdf

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumaan, H., and Hübner, S., 2001. Ontology -Based Integration of Information - A Survey of Existing Approaches, 17th *Joint Conference on Artificial Intelligence*, Seattle, Washington, USA (IJCAI-01) Workshop: Ontologies and Information Sharing, pp. 108-117