Knowledge-Based Systems 75 (2015) 104–123

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A semantic approach to data translation: A case study of environmental observations data

Yanfeng Shu [a,*], David Ratcliffe [b], Michael Compton [b], Geoffrey Squire [b], Kerry Taylor [b,c]

[a] CSIRO Digital Productivity, GPO Box 1538, Hobart, TAS 7001, Australia
[b] CSIRO Digital Productivity, GPO Box 664, Canberra, ACT 2601, Australia
[c] College of Engineering and Computer Science, Australian National University, North Road, Acton, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 19 April 2014
Received in revised form 28 July 2014
Accepted 20 November 2014
Available online 28 November 2014

Keywords:
Data translation
Declarative mapping
Ontology
Spreadsheet
XML-based exchange language

ABSTRACT

To facilitate the exchange of environmental observations, efforts have been made to develop standardised markup languages for describing and transmitting data from multiple sources. Along with this is often a need to translate data from different formats or vocabularies to these languages. In this paper, we focus on the problem of translating data encoded in spreadsheets to an XML-based standardised exchange language. We describe the issues with data that have to be resolved. We present a solution that relies on an ontology capturing semantic gaps between data and the target language. We show how to develop such an ontology and use it to mediate translation through a real scenario where water resources data have to be translated to a standard data transfer format. In particular, we provide declarative mapping formalisms for representing relationships between spreadsheets, ontologies, and XML schemas, and give algorithms for processing mappings. We have implemented our approach in AdHoc, an ontology-mediated spreadsheet-to-XML translation tool, and showed its effectiveness with real environmental observations data.

1. Introduction

Recent years have seen a proliferation of availability of environmental observations data as a result of the improvement of sensor technologies, the growing number, size, and complexity of environmental monitoring programs, and the realisation of the importance of using observations to characterise the environment as well as to describe it with models and simulations [1]. The successful use of these data to achieve new scientific breakthroughs, as well as in making well-informed resource management decisions, depends to a large extent on the ability to access, integrate and analyse these data [2].

Recognising this, there have been efforts to develop standardised markup languages for the exchange of environmental observations data. Examples include Water Data Transfer Format (WDTF) [3], Water Markup Language (WaterML) [4], Ecological Metadata Language (EML) [5], and Earth Science Markup Language (ESML) [6]. These languages provide a structured syntax for communicating data from multiple sources as eXtensible Markup Language (XML) documents. With each language, a set of specifications may be provided,[1] which describes additional requirements on data. Together, they define the information required, e.g. the location at which the observation was made, or the property that was observed; the constraints to be satisfied, e.g. "each observation has exactly one observed property", or "a water level must be measured in metres"; or the vocabulary terms to be used, e.g. "streamflow" instead of "flow" or "discharge".

On the other hand, there is no standardisation in the methods of data storage or management, and each data source can have its own methods for storing and managing its data. This gives rise to data with different formats or vocabularies. To access and use such data, there is often a need to translate them to a standardised exchange language (e.g. one of those mentioned above). For example, in Australia, in response to the increasing demand for improving the efficiency of water management practices, the Bureau of Meteorology (BoM) has been given a mandate to build and maintain an integrated national water information system, which involves collecting water resources data from over 200 organisations [7,8]. As the organisations involved use various software systems with many different data formats, WDTF was developed for

* Corresponding author.
E-mail address: Yanfeng.Shu@csiro.au (Y. Shu).

[1] In this paper, we consider a language and its specifications as a whole unless explicitly differentiated.

http://dx.doi.org/10.1016/j.knosys.2014.11.023
0950-7051/

transfer and ingestion of data into the national system. As a result, data from organisations have to be translated to WDTF; also, the translated data have to conform to a set of constraints and vocabularies [9].

The effort involved in data translation such as this[2] could be considerable. For *each* data source, it requires writing and managing complex data transformation programs or queries, including being familiar with the source data formats and vocabularies, and the target syntax and semantics (e.g. the constraints to be satisfied, and the vocabularies to be used). Although there are tools available to facilitate translation (by generating transformation queries), e.g. [11–14], these tools are designed for general use, focusing mainly on data with well-defined structure, and having little support for capturing the inherent meaning or semantics of data or the target language, and thus are insufficient for handling the type of translation discussed here, i.e. translation of environmental data from different formats or vocabularies to an XML-based standardised exchange language.

In this paper, we investigate ways to facilitate such translation. We focus on data in spreadsheets, as spreadsheets are commonly used to store environmental data. Starting by looking at some real data examples, we identify the issues with data that have to be resolved for data translation, including data being provided at various levels of information detail, having various structures, and using various terminologies and value representations. We then propose an *ontology-mediated* approach for data translation. We define an ontology for capturing semantic gaps between data and the target language; based on this, we use the ontology to mediate across different structures, terminologies and value representations of data, to check data against the constraints captured by the ontology and ensure that data be provided by the information required, and finally to produce data satisfying the requirements of the target language (in both syntax and semantics). The way the ontology is used makes it necessary to translate data into ontology instances first. We describe the approach in detail through the aforementioned water data translation scenario, including ontology development, spreadsheet to ontology mapping and translation, and ontology to XML mapping and translation. In particular, we provide declarative formalisms for mapping representation (to facilitate mapping customisation and reuse), and give algorithms for processing mappings.

To demonstrate the value of our approach, we have developed a tool (named AdHoc) for mapping construction and data translation. The tool provides a graphical interface for users to specify correspondences between spreadsheet data and the ontology. Based on correspondences, the system generates spreadsheet-to-ontology mappings, checks data constraints and translates data to the target XML format (all these are done in the back-end). Ontology-to-XML mappings are constructed manually, but only once, due to a single target language assumed in our work. We have applied AdHoc to the water data translation scenario, and showed its effectiveness with real water resources data. We note that because of the ontology-driven nature of AdHoc, it can be generally applied to other data translation scenarios that are not related to water, but have spreadsheets as data sources and XML as common exchange formats. In summary, we make the following contributions:

- We propose an ontology-mediated approach for environmental data translation, based on an analysis of the issues with data that have to be resolved.
- We outline the principles underlying the design of a mediating ontology for data translation, and show the development of such an ontology through a real environmental data translation scenario.
- We propose a declarative mapping formalism for representing the relationship between spreadsheets and ontologies, and give an algorithm for the evaluation of spreadsheet-to-ontology mappings.
- We propose a declarative mapping formalism for representing the relationship between ontologies and XML schemas, and give an algorithm for the evaluation of ontology-to-XML mappings.
- We have developed an ontology-mediated spreadsheet-to-XML translation tool, and showed its effectiveness with real environmental observations data.

The rest of the paper is organised as follows. In Section 2, we identify the issues with environmental data and discuss their implications on data translation. In Section 3, we present the proposed approach, and describe it in detail through the water data translation scenario. In Section 4, we report on the tool implementation and evaluation. Finally, we summarise related work and conclude the paper in Section 5 and Section 6, respectively.

2. Issues with environmental data

According to Beran and Piasecki [15], the biggest challenge in seamlessly integrating multiple data sources is resolving heterogeneity issues. This is also true when exchanging data from multiple sources, and translating data from different formats or vocabularies to standardised languages. Horsburgh et al. [2] classify heterogeneity in environmental observations data into two general types: syntactic and semantic heterogeneity. Syntactic heterogeneity refers to a difference in how data and metadata are organised (e.g. rows vs. columns) and encoded (e.g. text files vs. Excel spreadsheets). For this type of heterogeneity, we are mainly concerned about differences in data organisation or structure, as we assume in this paper that data are all encoded in spreadsheets.

Semantic heterogeneity, on the other hand, refers to the variety in language and terminology used to describe observations, including different languages used to describe the names of observation attributes, or to encode observation attribute values [2]. Semantic heterogeneity occurs when there is disagreement in the meaning, interpretation or intended use of the same or related data [16]. In the following, we illustrate both syntactic and semantic heterogeneity, and elaborate the issues involved that have to be resolved in data translation.

Fig. 1 shows 12 real data examples. Among these examples, (A) is about water usage information, (B) about ground water level information, (C) and (D) about watercourse level information, and (E)–(L) about water storage level or volume information. Although data in these examples are all encoded in spreadsheets, there is no fixed structure for data description; even for the same type of data, data structure could be different. For example, both (C) and (D) record watercourse level information; however, (C) stores water levels of one watercourse per spreadsheet, while (D) stores water levels of several watercourses per spreadsheet. As another example of structural differences, water storage names in (K), (L) and (E) are listed as column names, while in (I) they are stored as column values.

Besides differences in data structure, the examples in Fig. 1 also expose several semantic heterogeneity issues. One is that contextual information or metadata is provided at various levels of detail. Contextual information is the descriptive information about data that explains the measurement attributes, their names, units, precision, accuracy and data layout, as well as the data lineage describing how the data was measured, acquired, or computed

[2] There are similar cases in other domains as well. As XML becomes a common standard for data exchange, legacy data are often required to be placed into a predefined XML schema (defined, e.g. by a standards committee to permit meaningful exchange within a specific domain) [10].