ELSEVIER

Computers & Geosciences 31 (2005) 1126–1134

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Syntactic and semantic metadata integration for science data use

Sunil Movva, Rahul Ramachandran*, Xiang Li, Sarita Khaire, Ken Keiser, Helen Conover, Sara Graves

Information Technology and Systems Center, University of Alabama in Huntsville, Huntsville, AL-38599, USA

Received 14 May 2004; received in revised form 6 October 2004; accepted 31 December 2004

Abstract

This paper proposes a novel metadata solution to allow applications to intelligently use science data in an automated fashion. The solution provides rich syntactic and semantic metadata, where the semantic metadata is linked with an ontology to define the semantic terms. This solution allows applications to exploit the syntactic metadata to read the data and the semantic metadata to infer the content and the meaning of the data. The solution presented in this paper leverages the Earth Science Markup Language for providing the syntactic metadata and adds a semantic metadata component along with links to the appropriate ontology. This new semantic component is orthogonal to the syntactic metadata, so it does not perturb the existing design. An example application was designed and built that integrates this syntactic and semantic metadata via an ontology to perform a data processing operation.

Keywords: XML; Metadata; ESML; Ontology; DAML + OIL

1. Introduction

Metadata, or data about data, for science data sets can be classified into three general categories: content, syntactic and semantic. Content or "search" metadata is a broad category describing the intrinsic content of the data. *Content metadata* typically describe the physical parameters or variables measured in a data set, its spatio-temporal coverage, coordinate systems used, information about the data producer, provenance and other keywords. Content metadata typically populate data catalogs and registries. *Syntactic metadata* describe the structure of the data file in terms of bits, bytes, data type, arrays and structures. This information is often found in README files accompanying science data. For some data formats, this information is embedded in the data file and an accompanying software library is used to create or read the files in these formats. Finally, *semantic* or "use" metadata provide meaning to the data, relating the content of the data file to some known context. Such semantic information may be found in documentation or publications about a data set. Current semantic web research is aimed at encoding semantic information in ontologies in order to enable more powerful and intelligent automated data search and usage.

To accommodate the rapid growth of Earth Science data, scientific investigations require the ability to use new data sets with minimal effort. With the emergence of Semantic Web (Berners-Lee et al., 2001) concepts,

*Corresponding author. Fax: +1 256 824 5149.
E-mail address: rramachandran@itsc.uah.edu (R. Ramachandran).

0098-3004/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2004.12.011

intelligently automating services has become an area of active research. Extending this idea of intelligently automating services to the Earth Science domain requires solving several critical issues related to the varying level of metadata richness associated with different geosciences data sets. This paper proposes a metadata solution by integrating syntactic with semantic metadata to automate data use. This solution assembles orthogonal and yet synergistic information to provide a rich description of the syntax and semantics of scientific data sets. Furthermore, data analysis and other applications are provided a context for the semantic metadata via an ontology. Applications can use this rich set of metadata with the ontology to make automated decisions to achieve data processing goals. This paper describes an example application that uses this metadata information to automate and drive a useful data preprocessing capability.

The concept of integrating syntactic and semantic metadata for data use is not new (Cornillon et al., 2003). However, the solution presented in this paper is the first of its kind in Earth Science: the metadata contains both syntactic and semantic information; the semantic metadata is linked to an ontology to provide context; and an application is designed and built to use this metadata and the ontology to perform data processing.

2. Syntactic and semantic metadata integration issues

Large quantities of raw and processed observational data and imagery are available to researchers today. These data sets are heterogeneous and with varying levels of metadata richness. It is essential that a syntactic and semantic metadata integration solution take into account existing legacy data sets. In order to design such a solution, one must address the following issues.

2.1. Read data formats with varying levels of syntactic metadata

In order to use the science data, applications must be able to read different formats. Within the Earth Science communities, many data products are being produced using self-describing data formats, which contain varying levels of metadata. There are also many data formats (especially legacy data) that contain no metadata at all. Thus, Earth Science data formats can be categorized as "free" or "self-describing" data formats. Data files in Binary and ASCII text encoding, lacking syntactic metadata, are considered "free" formats. These data files do not contain sufficient or any metadata information to allow development of a general software tool that can extract data fields from these files. Also, the same data can be structured in several different ways depending upon the data producer's preference. 'Self-describing' data formats such as HDF, netCDF, or HDF-EOS do contain syntactic metadata. However, software applications have to interface with a separate library for each of these data formats to extract data. A metadata solution must provide syntactic metadata for data in free formats without requiring laborious data translations, and common interfaces to both free and existing structured data formats. Recently designed data formats can package data and metadata together using a standardized XML schema (Shaya, 2002; Williams, 2000). However, a pure XML solution is not practical because of the complexity and volume of some of these geospatial data products. Moreover, a solution that relies on a single standard data format or a Markup Language that does not consider the existing (legacy) data resources and other metadata efforts will not provide the flexibility needed by the science community. Given the enormous volumes of data that have already been archived and cataloged by agencies such as NOAA, NASA and USGS, any solution that requires reformatting data products or duplicating databases would be extremely expensive and possibly counterproductive.

2.2. Provide semantic metadata with context

Structured data formats allow data users to specify fields within a data file by name. However, descriptive field names cannot be used by applications to automate processing without context information. For example, if a data field is named "rainfall_rate", without the context knowledge that "rainfall rate" is a subsumption of "precipitation", an application will not be able to address a user query to extract all precipitation fields from the data. A metadata solution is required that allows annotating data fields with semantic metadata, where the context of the semantic metadata are defined in an ontology.

2.3. Provide a robust semantic metadata solution

Initial attempts at providing semantic metadata descriptions for data fields have often been problematic. This is because the meaning of the metadata is hard coded into the software application, making this solution brittle. For example, the HDF-EOS format allows the data producer to identify geolocation fields using the names "Latitude" and "Longitude". Thus, any field named "Latitude" can then be used by the HDF-EOS software library to navigate the data and to perform spatial subsetting. However, in many instances of data sets created in HDF-EOS, the latitude data field was labeled with a variant name such as "latitude" or "lat". These data sets were not able to utilize the geographic subsetting functionality of the software library. Semantic metadata solutions should not rely on hard coded applications, but should be flexible