Data-Intensive Scientific Management, Analysis and Visualization

Mariana Goranova, Bogdan Shishedjiev, Juliana Georgieva

Technical University of Sofia, Sofia, Bulgaria

E mail mgor@tu-sofia.bg

Accepted: December 1, 2011

Abstract. The proposed integrated system provides a suite of services for data-intensive sciences that enables scientists to describe, manage, analyze and visualize data from experiments and numerical simulations in distributed and heterogeneous environment. This paper describes the advisor and the converter services and presents an example from the monitoring of the slant column content of atmospheric minor gases..

Keywords: eScience, forth paradigm, data-driven approach, service-oriented architecture, database management system

Introduction

Scientific communication produces massive volumes of unstructured and heterogeneous data due to the various data sources. The data analysis explores interesting patterns that discover new theories. The new "fourth paradigm" (Gray, 2007; Hey, Tansley and Tolle, 2009) focuses on data-intensive science (eScience) that uses the data-driven approach and requires a collaboration between the scientific and computer science communities. The basic activities of science are modified during the last years from empirical, theoretical and computational branch through data-intensive science that consists of data capture, curation and analysis. The data-intensive science needs new ways to manage the massive amounts of data, captured by instruments and simulations, integration of software analysis tools directly into the database, support of visualization and interactivity, sharing the data sets among researchers.

Addressing the challenges of the data-intensive science, our investigation is in the area of "scientists' smart environment" – methods on scientific data management and visualization of data from experiments and numerical simulations in distributed and heterogeneous environment, the necessary tools and architectures. We use the paradigm of service-oriented architecture (SOA) that provides the ability to locate and invoke a service across machine and organisational boundaries, both in a synchronous and an asynchronous manner. The SOA-based approach (Erl, 2005; Rosen et al., 2008) for processing, querying, accessing, and retrieving data provides functionality to scientists easily to capture, organize, analyze, discover, visualize and publish data. Scientists access the system over the web using services and they are able to flexibly orchestrate these services into computational workflows.

The science community develops a lot of data formats used for data interchange and carrying out the data schema. Most of these formats are XML-based, related with specific and concrete data and do not have semantic description of data. Hierarchical Data Format (HDF, 2009), Data Format Description Language (DFDL, 2008), Earth Sciences Markup Language (ESML, 2006) are examples of XML languages that deal with data of specific domains. Different disciplines of science need much better tools that integrate all the structure and the semantics of data and can deal with data from measurements and simulations.

This paper is organized as follows. We first briefly review the architecture of the proposed integrated system and the developed XML-based language for specific scientific data description. We talk about the services that implement raw data description, automatic conversion of scientific data into canonical format and then we present data examples from daily ground-base spectrometric measurements with the help of the GASCOD-BG instrument. Lastly we draw our future work and conclusions.

Data management

Most scientific data is unstructured and heterogeneous due to the various data sources. Different data formats need meta information about structure and usage. The essential aspects of efficient data management are semantic annotation, location, consistency maintenance and sharing between multiple users in distributed data systems.

Architecture for scientific data management

"Service-oriented architecture, or SOA, is the modern notion of connecting systems together at both the information and service levels" (Linthicum, 2004). SOA is a model in which automation logic is decomposed into smaller, distinct units of logic that can be distributed. The large, monolithic application is breaking up into separate components (called services), that use standardized connections to each application. The enterprise-related motivation includes the ability to readily change business processes on top of existing services and information flows and to monitor points of information and points of service in real time. Developers can choose the right enabling technology (C#, Java, Cobol, C++) for the job without

concerning themselves with technical dependencies and security models.

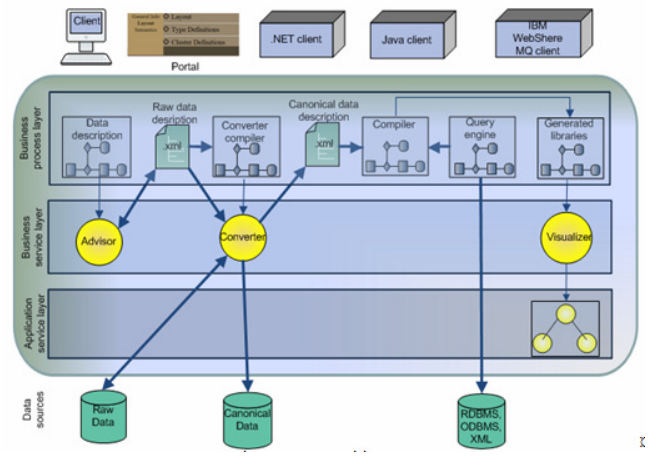

Fig. 1. SOA architecture

Figure 1 shows the proposed system architecture (Goranova et al., 2009) that enables the linkage of data and resources and provides functionality to scientists, who access the system over the Web. The architecture allows the data captured by instruments or generated by simulations to be processed and visualized. The physical scientist uses the **Advisor** service to describe the data and the resulting XML raw data description is stored in a database while the large quantities of raw data are stored on files servers. The **Converter** service automatically transforms the data from the original layout to another one based on the semantics named canonical format providing a link between raw representation of data and database schemas. The resulting canonical data description translates the scheme from the XML model into a form that looks like a relational data model. The **Visualizer** service retrieves and queries data from data analysis and visualization tools. This loosely coupled system allows datasets to be widely available via Internet.

XML-based description language

We have developed a definition of language (Shishedjiev, Goranova and Georgieva, 2010) that integrates both the structure and semantics of data without using a separate ontological language for semantic description. Our goal is to use the proposed language for modeling and simulations of experimental data from various sources – from radiation and spectral measurements made in scientific institutes of the Bulgarian Academy of Sciences (BAS) and simulation data for magnetic fields used in medicine.

The developed language describes the structure and semantics of scientific data set based on the XML schema. Every XML document consists of root element **`<dataset>`** with three base elements: **`<general>`**, **`<semantics>`** and **`<layout>`**.

The <**general**> element (Fig. 2.a) defines and annotates data. It contains elements describing the general characteristics of the data set as identification, annotation, ownership, permissions, version history, procedure creation, etc. It consists of required and optional sub-elements.

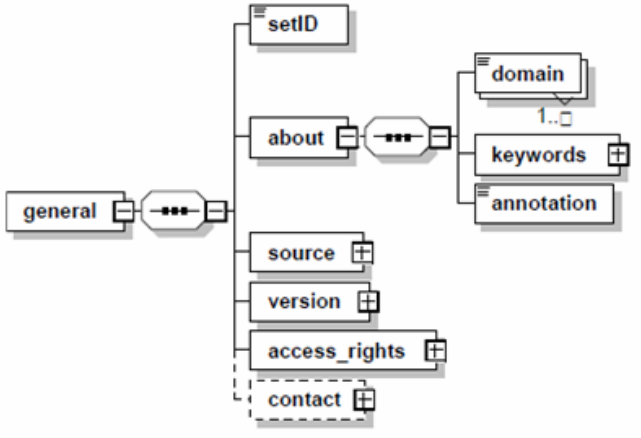

Fig. 2.a. <general> element

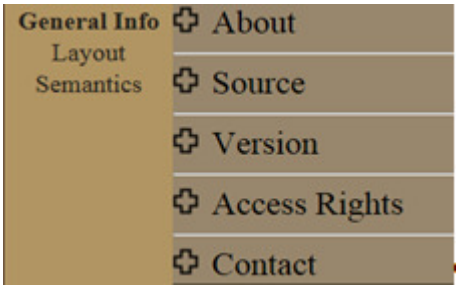

Fig. 2.b. General form

The <**semantics**> element (Fig. 3.a) describes the meaning of data. The sub-element <**data_parameters**> represents the group of parameters related to the entire data set. The <**independent**> sub-element includes the most popular quantities as <**time**> and <**space**>; <**other**> is any possible other type of independent parameter. The <**dependent**> contains <**field_value**> element and <**dependent_on**> element that contains dependency information. All semantic elements contain the <**layout_ref**> element that is used to indicate the corresponding item in section <**layout**>. The <**semantics**> part is the base for the canonical structure of the data because our assumption is that data may be presented as a table containing the values of independent and dependent variables.

The <**layout**> section (Fig. 4.a) contains two sub-elements: <**typedef**> and <cluster>. The <**typedef**> element contains user defined types as types in programming languages with additional restrictions. The dataset can contain a lot of files and the structure of each file is described by one <**cluster**> sub-element. The base <**item**> element can be simple (<**simple_item**>) or complex (<**structure**> and <**array**>) and contains the <**semantics_ref**> element as a reference to the corresponding part in <**semantics**> section.

Implemented services

1.Advisor

The **Advisor** helps the scientists to translate their data into XML format. It allows scientists to describe

their specific file formats (binary or text) and writes the data in file descriptors following the rules specified in the schema of the developed XML-based description language. The graphical user interface developed with ASP .NET technology aims to make the description of scientific data easy and intuitive so that every user can work with the Web based solution.

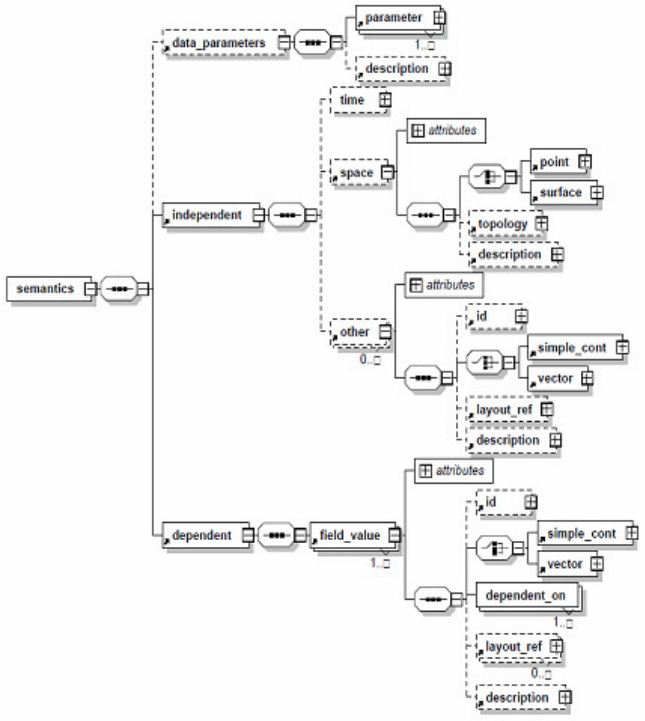

Fig. 3.a. <semantrics> element

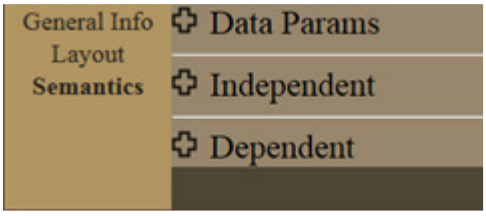

Fig. 3.b. Semantic form

The **Advisor** contains three forms – **General Info**, **Semantics** and **Layout**.

The general form (Fig. 2.b) contains information identified the dataset, the domain and the data source – how data is obtained, the used procedures and instruments, methods to process data. It also includes date and time of creation, version number and the features. Access rights are divided into three groups: ownership data, distribution data and read data. The general form includes information about the responsible organization and the contact person.

The semantic form (Fig. 3.b) consists of three forms. **Data Params** describes parameters valid for the data set as a whole that cannot be referred to independent nor dependent component. **Independent** form contains the independent variables as time and space and all other parameters. **Dependent** form includes all dependent variables from the dataset.

The layout form (Fig. 4.b) describes the data layout in the dataset, the internal representation of data, the type of encoding (ASCII, Binary, Unicode, EBCDIC) and the byte order (Big-endian or Little-endian). This form includes also the type definition of complex objects as structures, arrays and simple items.

The generated XML raw data description file is stored in the database and is transmitted to the converter.

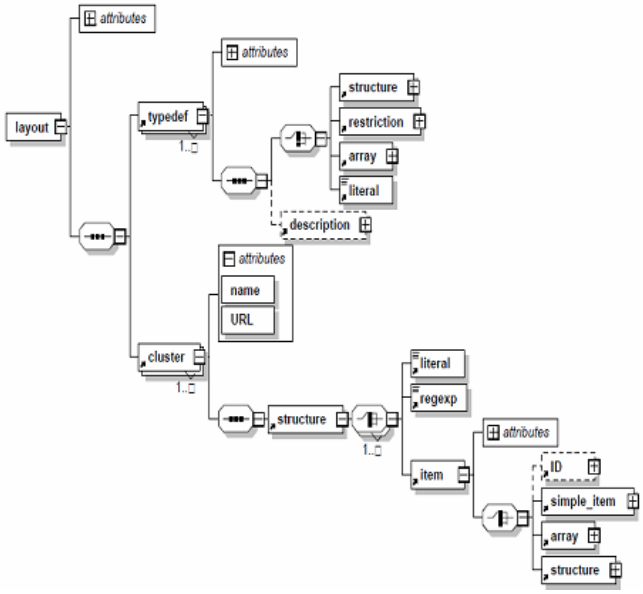

Fig. 4.a. **<layout>** element

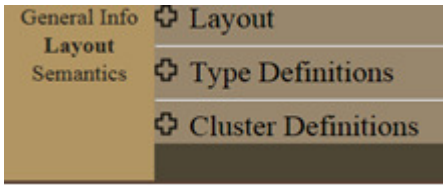

Fig. 4.b. Layout form

2.Converter service

Figure 5 shows the canonical format grammar presented as XML-Schema. The canonical format has a root **<canon>** element consisting of **<table>** elements. The **source** attribute of **<table>** contains the address of the XML raw data description file that is used as the input of the converter compiler (Goranova et al., 2011).

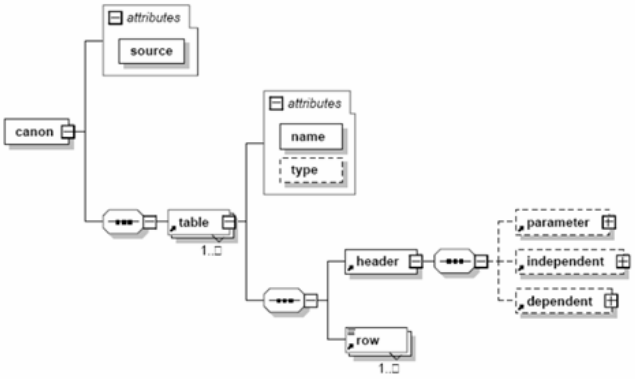

Fig. 5. Canonical format grammar presented as XML-Schema

Converter compiler using the **<semantics>** section creates a collection of columns that construct the **<header>** sub-element of each **<table>**. Each column has a **name**, **type** and **reference** to columns of other tables to represent the relationship between the corresponding tables. The **<raw>** sub-element represents the values of the table rows. The Converter compiler reads the **<layout>** section of the raw data description file and creates a collection of objects representing a single structure entity in the data description and fills the **<raw>** elements of the table. The relation between the structure of the canonical form and the read data is ensured. The resulting canonical XML description file has a structure that looks like a relational data model and is stored in the database.

Experimental results

We present an example based on the O_3 and NO_2 measurements that are important for construction of adequate climate model. The monitoring of the slant column content of such atmospheric minor gases is very important for climate change problems and environment protection. These spectrometric measurements are carried out at Stara Zagora department of Space and Solar Terrestrial Research Institute – Bulgarian Academy of Sciences (SSTRI – BAS) and are obtained by the GASCOD-BG (Gas Analyser Spectrometer Correlating Optical Differences) instrument. The objectives of the data analysis are to find out the variability of the NO_2 and O_3 and its reasons and the establishment of long time trends is very important with regards to the climate change.

Figure 6 shows the raw data file structure that contains the sequence of spectral measurements for each day (Werner, 2008). The GASCOD-BG instrument operates in automatic mode but the measurement duty cycles are sometimes interrupted and the data file contains undefined number of blocks. Each block begins with a header with additional technical information: date and time of the measurement, the wave length (lambda), filters ($filter_1$ and $filter_2$) and integration time. The integration time is limited between 2500 and 239900 seconds where the value 2500 means that the instrument is in the calibrated state. Each spectrum contains 512 values (pixels). Spectral data are grouped in two dimensional matrixes of 51 rows and 10 columns plus one 52nd row with the last two values.

The resulting XML raw data and canonical data descriptions are shown in Fig. 7 and Fig. 8, respectively. The raw data description describes the structure, semantic and annotation of scientific data. The canonical form groups the data into logical blocks looking like tables and realizes the principles of the relational data model.

Our future work will continue with creating a query engine to access and query data. The next step is developing and applying new numeric and visual models and methods to visualise data.

Our work considers the opportunities and challenges for data-intensive science and establishes cooperation among scientists of SSTRI – BAS and computer specialists of the Technical University of Sofia. We follow the current trends in the field of eScience discusses in Hey, Tansley and Tolle, 2009.

Conclusions

Our contributions are in the nontrivial process of describing of observational data because of various data formats which requires sophisticated techniques. Understanding of observational data begins by understanding the information regarding the origins, ownership, metadata, and structure layout of datasets. Our developed data model automatically converts the raw data description into canonical form with a search engine that doesn't require advanced knowledge of database management. The visualization with highly interactive possibilities is an essential part of exploration and analysis.

The proposed data-intensive scientific system permits Web based access allowing scientists to explore and gain better understanding of natural world but limited technical background in data management and analysis to explore large quantities of data.

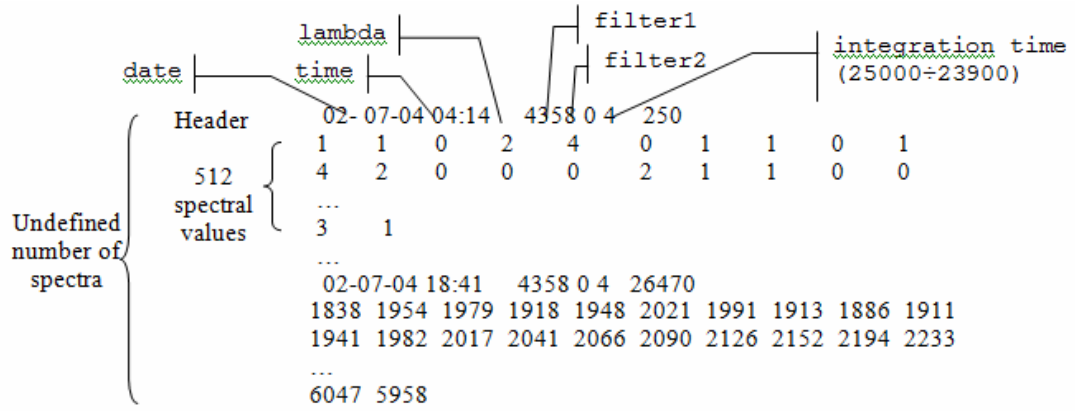

Fig. 6. Data file structure

```
<dataset>
 <general>
  <setID>Spectra_II_2234/>
  <about>
   <domain>Measurements of O3 and NO2/>
   <keywords>…</keywords>
   <annotation>…</annotation></about>
  <source>
   <timestamp>2004-07-04T23:00:00/>
   <method>
    <instrument>GASCOD-BG/>
   </method></source>
  <version>…</version>
  <access_rights><ownership>
   <Name><firstName>Rolf/>
         <familyName>Werner/></Name>
   <persAddress>
    <country>Bulgaria</country>
    <city>Stara Zagora</city>
    <email>werner@yahoo.co.uk</email>
   </persAddress></ownership> </access_rights>
  <contact>
   <ContactPerson>…</ContactPerson>
   <ContactOrganisation>…
   </ContactOrganisation></contact>
 </general>
 <semantics>
  <independent>
   <time>
    <component name="date">…</component>
    <component name="hour">
     <parameter name="lambda">…</parameter>
     <parameter name="filter1">…</parameter>
     <parameter name="filter2">…</parameter>
     <parameter name="integ_time">…
     </parameter>
    </component></time>
   <other name="pixnom">…</other>
  </independent>
  <dependent>
   <field_value name="measure_point">…
   </field_value>
  </dependent>
 </semantics>
```

```
 <layout coding="ASCII">
  <typedef typename="measurement_value">…
  </typedef>
  <typedef typename="header">
   <structure>
    <item name="date_meas">…</item>
    <literal>" "</literal>
    <item name="time_meas">…</item>
    <literal>"     "</literal>
    <item name="lambda">…</item>
    <literal>" "</literal>
    <item name="filter1">…</item>
    <literal>" "</literal>
    <item name="filter2">…</item>
    <literal>"  "</literal>
    <item name="integration_time">…
    </item></structure></typedef>
  <typedef typename="measurements">
   <array length="512">
    <item name="point">…</item>
    <separator>
     <regexp>[\s,\r?\n]*</regexp>
    </separator></array></typedef>
  <typedef typename="record_meas">
   <structure>
    <item name="measurement_header">…
    </item>
    <literal>EOL</literal>
    <item name="measurement">…</item>
   </structure></typedef>
  <cluster name="cluster" URL=…>
   <structure><item type="array"
                    name="spectral_data">
     <array>
      <item name="meas_line">…</item>
      <separator><literal>EOL</literal>
      </separator>
      <end_criteria>
       <literal>EOF</literal>
      </end_criteria></array></item>
   </structure>
  </cluster>
 </layout>
</dataset>
```

Fig. 7. Raw data description

```
<canon>
<table name="param">
 <header>
  <independent>
   <time>
    <component name="date">…</component>
    <component name="hour">
     <parameter name="lambda">…</parameter>
     <parameter name="filter1">…</parameter>
     <parameter name="filter2">…</parameter>
     <parameter name="integ_time">…
     </parameter></component></time>
  </independent>
 </header>
 <row>02-07-04 04:14 4358 0 4 2500</row>
 <row>…</row>
 <row>02-07-04 18:41 4358 0 4 26470</row>
</table>
```

```
<table name="base">
 <header>
  <independent>
   <time>
    <component name="date">…</component>
    <component name="hour"></component>
   </time></independent>
  <dependent>
   <field_value name="measure_point">…
   </field_value></dependent></header>
 <row>02-07-04 04:14 1 1</row>
 <row>…</row>
 <row>02-07-04 04:14 512 1</row>
 <row>02-07-04 18:41 1 1838</row>
 <row>…</row>
 <row>02-07-04 18:41 512 5958</row>
</table>
</canon>
```

Fig. 8. Canonical data description

Acknowledgements

This work was funded by the National Scientific Fund of Ministry of Education and Science in Bulgaria (Grant Number DO 02-175/2008).

References

DFDL (Data Format Description Language): 2008, http://forge.gridforum.org/projects/dfdl-wg.

ESML (Earth Sciences Markup Lanaguage): 2006, http://esml.itsc.uah.edu.

Erl, T.: 2005, SOA: Concepts, Technology, and Design, Prentice Hall PTR, ISBN: 0-13-185858-0, p. 58.

Goranova, M., Shishedjiev, B., Georgieva, J., and Todorova, V.: 2009, Architecture for processing, managing and visualisation scientific data in SOA environment, in Proceeding of the Seventh International Conference on Challenges in Higher Education and Research in the 21st Century, Sozopol, Bulgaria, vol. 7, p. 309.

Goranova, M., Shishedjiev, B., Georgieva, J., and Achev, V.: 2011, Automatic Conversion of Scientific Data into Canonical Format, in Proceeding of EUROCON – International Conference on Computer as a Tool, Lisbon, Portugal, IEEE Computer Society, p. 1, ISBN: 978-1-4244-7486-8, DOI: 10.1109/EUROCON.2011.5929171.

Gray, J.: 2007, eScience – A Transformed Scientific Method, eScience Talk at NRC-CSTB meeting, Mountain View CA, http://research.microsoft.com/en-us/um/people/gray/JimGrayTalks.htm.

Hey, T., Tansley, S., and Tolle, K.: 2009, The Fourth Paradigm Data Intensive eScience Discovery, Microsoft Corporation, p. 4.

HDF (Hierarchical Data Format): 2009, http://www.hdfgroup.org/HDF5/.

Linthicum, D.: 2004, Extending Your SOA for Intercompany Integration, http://www.virtualizationconference.com/node/45096.

Rosen, M, Lublinsky, B., Smith, K, and Balcer, M.: 2008, Applied SOA. Service-oriented Architecture and Design Strategies, Wiley Publishing, Inc., p. 246, ISBN: 978-0-470-22365-9.

Shishedjiev, B., Goranova, M., Georgieva, J.: 2010, XML-based Language for Specific Scientific Data Description, in Proceedings of the Fifth International Conference on Internet and Web Applications and Services, ICIW 2010, Barcelona, Spain, IEEE Computer Society, p. 345, ISBN: 978-0-7695-4022-1.

Werner, R.: 2008, The Latitudinal Ozone Variability Study using Wavelet Analysis, Journal of Atmospheric and Solar-Terrestrial Physics, Volume 70, Issue 2-4, p. 261.