Knowledge Consortium of Gujarat
Department of Higher Education - Government of Gujarat
Journal of Science
ISSN: 2320 - 0006

Continuous issue-5 | January – March 2016

Analysis of Web technologies for environmental Big Data

Abstract

Recent evolutions in computing science and web technology provide the environmental community with continuously expanding resources for data collection and analysis that pose unprecedented challenges to the design of analysis methods, workflows, and interaction with data sets. In the light of the recent UK Research Council funded Environmental Virtual Observatory pilot project, this paper gives an overview of currently available implementations related to web-based technologies for processing large and heterogeneous datasets and discuss their relevance within the context of environmental data processing, simulation and prediction. We found that, the processing of the simple datasets used in the pilot proved to be relatively straightforward using a combination of R, RPy2, PyWPS and PostgreSQL. However, the use of NoSQL databases and more versatile frameworks such as OGC standard based implementations may provide a wider and more flexible set of features that particularly facilitate working with larger volumes and more heterogeneous data sources.

1. Introduction

1.1. The internet for sharing and linking data and models

Environmental sciences are witnessing a rapid increase in the amount of relevant published information on the internet. A large share of these data is the result of environmental monitoring; either in situ or via remote sensing that is being made available by government institutions, private companies and citizen scientists. However, many other data that are not collected for environmental purposes may be useful for environmental science. Examples include geotagged photographs that may contain information about land cover and hydro meteorological conditions, disturbance patterns in telecommunication systems that provide information about weather patterns, data feeds from internet-enabled objects, online social network interactions, and many others.

1.2. Big Data

Big Data is defined as any collection of data sets which volume and complexity make data management and processing difficult to perform using traditional tools (i.e. handling N-dimensional data sets using plain text files and/or SQL databases). Those problems invest Big Data monoliths as much as ecosystems of small data causing major concern for most private and public data providers for which “small quantities do not equal simpler management”. Even though Big Data is usually associated with the LOD concept, it is generally comprised of linked and non-linked data, open and private data, and, as such, it is characterized as being composed of the “three Vs”: significant growth in the volume, velocity and variety of data. In this review, we include in the above definition of Big Data also the collection of technologies that cope with the effects of this abundance and heterogeneity, proposing solutions to meet the needs of a modern scientific community.

1.3. Resource and approaches for using Big Data

Although new desktop applications are being developed to provide a number of Big Data analysis tools, the internet, as well as being a source of data, also provides powerful tools for data processing, visualization, simulation, prediction and sharing. A variety of projects in different countries are analyzing how this potential can be harnessed, such as the UK Natural Environment Research Council funded Environmental Virtual Observatory pilot project3, the Earth Cube initiative of the US National Science Foundation4, and the Global Earth Observation System of Systems.

Because of the reliance on standardized data exchange, the internet provides a powerful environment to orchestrate complex workflows that rely upon distributed and modular components, chained together by web service technologies, as suggested by Dietze et al. who proposed the paradigm of "models as scaffold" to integrate data sources and data sets on different spatial/temporal/organizational scales.

1.4. Motivations and outline

The purpose of this paper is to introduce the range of available tools and technologies for web-based environmental modeling but also document investigations undertaken by the authors when prototyping the EVOp. EVOp's aim was to link data, models and expert knowledge to make environmental monitoring and decision making more efficient and transparent to the whole community. Therefore robust and reproducible methods to access and manipulate available data were needed along with effective communication tools tailored to be used by users with different levels of expertise.

This paper, therefore, reviews the current state of art of web based environmental data processing tools in the Big Data era. We believe this is of great significance to a myriad of efforts within the different scientific communities that are aiming to capitalize on such tools to build research collaboration environments and virtual observatories. The paper particularly focuses on the technological advancements and standards that are relevant to the environmental science community. We shed light on the different options available for assembling web service architectures, detailing aspects pertaining to data management and manipulation. We include examples of efforts that relate to the subject, describing the technological contribution of each specific project.

2. Web services and system architectures

Web services are essential in the orchestration of internet-based workflows. In essence, a web service is an application that enables access to its functions using established internet standards. As such they provide seamless cross-platform interoperability between different loosely coupled systems. Currently, two main architectural styles are most commonly used: SOAP and REST.

SOAP services use remote procedure calls to invoke functions on remote systems. Means of invocation (i.e. functions, parameters, return values, etc.) are described using the Web Services Description Language (WSDL). Using SOAP, clients generate "stubs" to match the service's interface. Data sent over the network is serialized into a structured XML format, which makes it machine-readable and implementation-independent. SOAP services can discover more services through a UDDI registry (similar to a directory service), while users can do so through data portals with search capability. This architecture relies on a host of further specifications to govern such aspects as security, privacy, and reliability of message exchange.

In a SOAP web service, instead, parameters are passed through a POST payload, as in the example below.

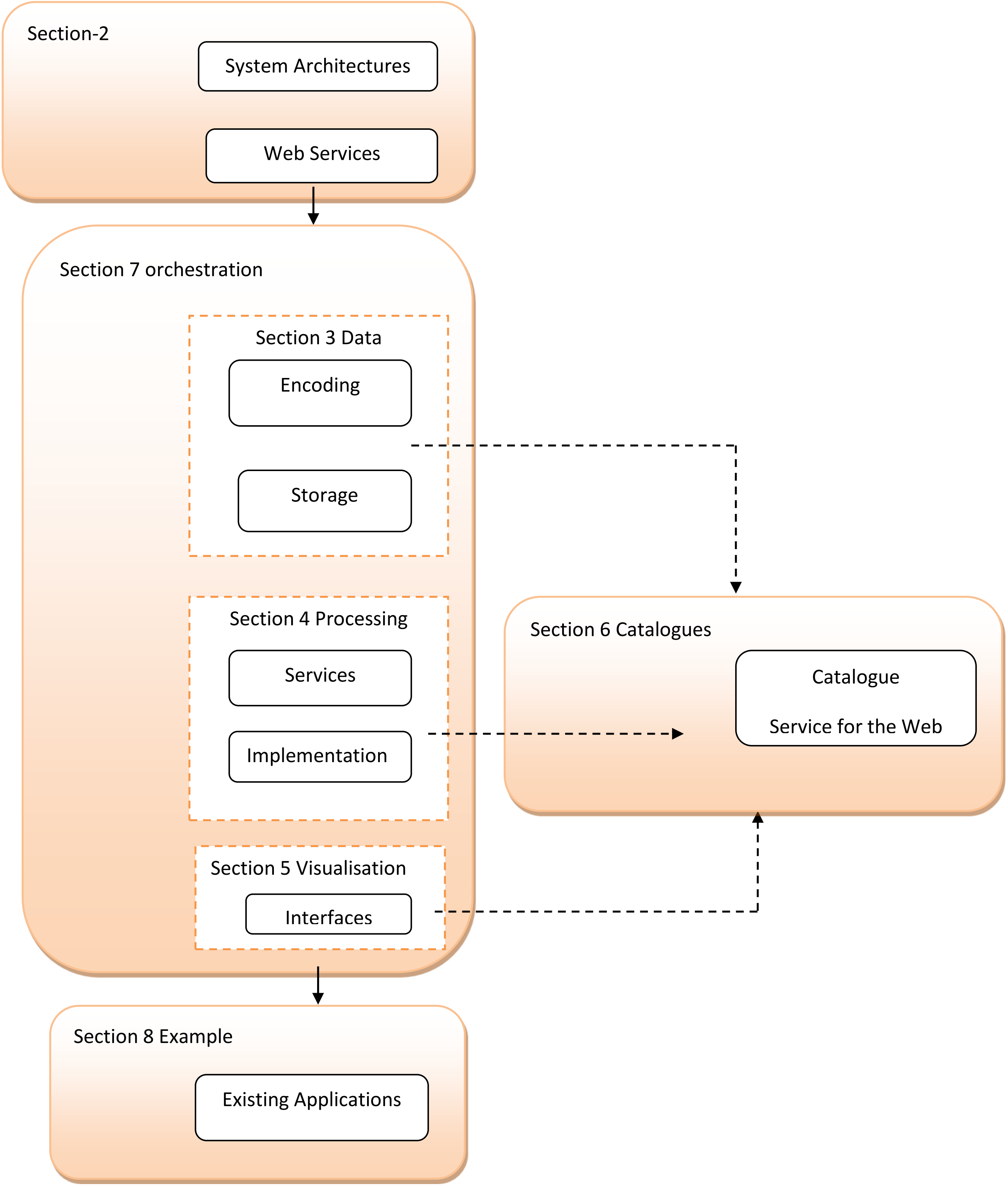
```
POST www.server.com

Path: /pywps/pywps.cgi

<?xml version="1.0"?>

<soap:Envelope

        xmlns:"http://www.w3.org/2001/12/soap-encoding">

<soap:Body>

<m:RunService>

        <m:service>wps</m:service>

        <m:request>execute</m:identifier>

        <m:identifier>mymodel</m:identifer>

</m:RunService>

</soap:Body>

</soap:Envelope>
```

3. Data standards

3.1. Data encoding

At the upstream end, an environmental data processing workflow typically starts with one or several datasets. In a web environment, relevant datasets are retrieved from data services available either locally or over the internet. Depending on the service and the type of information, data can be presented in different formats. Modeling platforms are, therefore, required to interact with a mixture of data formats, including plain text, markup languages and binary files. To enable cross-client and cross-platform compatibility, some currently existing web-based data services adopt a plain text format. The Critical Zone Observatories10 and the Geoinformatics for Geochemistry System11 are examples of database web services adopting plain text format. Their integrated systems store data, whenever possible, as an ASCII text table. The attached metadata uses an expanded Observations Data Model vocabulary or a unique sample identification code to retrieve data and set standards for metadata and data reporting. A user can also retrieve data manually, as these services make available map interfaces and visualization tools along with analysis tools.

A main advantage of using plain text is its accessibility without specific tools, which makes the method future-proof. However, from the viewpoint of workflow orchestration, extracting information is much easier if the format of the plain text file is self-describing. This can be achieved using a markup language which is not intended to be human-readable but machine parsable. The eXtensible Markup Language (XML) is a common standard for data interchange. Utilizing XML has many advantages: it combines data and metadata in one single file, it uses a text format and complies with well documented standards. As a cross-platform format, it is not exclusive to any particular operating system or development platform and it is typically well supported by data management software such as databases and GIS platforms. Additionally, specific varieties of XML have been developed for handling environmental data.



3.2. Data provenance

In the context of semantic web services (SWS), data provenance is becoming increasingly important for inspecting quality, usability and reliability of data in distributed computing environments. Behind the concept of provenance is the dynamic nature of data. Data captured and/or archived for environmental purposes continues to evolve over time as it is transformed and analyzed through different tools and by different organizations.

3.3. Data storage

Relational databases were first introduced by Codd (1970) and are currently the predominant choice in storing and sharing environmental data. A common Relational Database Management

System (RDBMS) assumes that data can be organised in tables (with a relatively simple structure) and that relations set among tables can be used to perform complex queries. Some popular RDBMS options are: PostgreSQL and MySQL. They both use standards such as SQL and XML and can therefore support data formats mentioned in the previous section. Technologies to handle explicitly spatial data are also well established, with specific data schemas and high performance processing options for their large file sizes and specific structures.

Conclusions

This paper presents a review of the most relevant web technologies dealing with “Big Environmental Data”. A common thread and the main motivation of this work is to document investigations carried out when prototyping the UK Environmental Virtual Observatory pilot. Evidence has shown that technologies can be effectively combined in many different ways depending on the specific modeling needs. However domain-specific projects require often tailored solutions. Numerous options for data formats, storage, processing, visualization and chaining of service components are taken into consideration. We found that, for example, despite the common practice of using plain text, self-describing data formats would be a better solution to store and transfer environmental data as they could integrate metadata information and standardized definitions of domain-specific variables and uncertainties. Also, as larger volumes

of data become available, data becomes less structured and therefore more complex. NoSQL databases have been found to deal better with complex and non structured information than traditional relational databases. Even though web-based processing can be approached in many different ways, at the moment the 52North framework seems to provide the most comprehensive and well supported platform currently available. Javascript libraries provide great potential to enable highly customized and interactive web-based visualization. A clear separation line cannot be drawn, instead, for the most popular workflow orchestration tools, which functionalities are very similar.

References:

1) Akers, K.G., Feb. 2013. Looking out for the little guy: small data curation. Bull. Am.Soc. Inf. Sci. Technol. 39 (3), 58e59.
2) Ames, D.P., Michaelis, C., Anselmo, A., Chen, L., Dunsford, H., 2008. MapWindow GIS.
3) In: Shekhar, S., Xiong, H. (Eds.), Encyclopedia of GIS. Springer US, Boston, MA,pp. 633e634.
4) Ames, D.P., Horsburgh, J.S., Cao, Y., Kadlec, J., Whiteaker, T., Valentine, D., Nov. 2012.HydroDesktop: web services-based software for hydrologic data discovery, download, visualization, and analysis. Environ. Model. Softw. 37, 146e156.

****

Dr. Pratik A Vanjara
Department of Computer Science
Shree M. & N. Virani Science College
Rajkot