GeS 2007 German e-Science

Available online at http://www.ges2007.de

Continuous digital workflows for earth science research

J. Klump, P. Löwe, R. Häner and J. Wächter

Data Centre, GeoForschungsZentrum Potsdam,
Telegrafenberg, 14473, Potsdam, Germany
Email: {jens.klump, peter.loewe, rainer.haener,
joachim.waechter}@gfz-potsdam.de
phone: (+49 0331) 288 1702, *fax:* (+49 0331) 288 1703

Abstract

The wealth of data available in the earth sciences is underutilised due to the absence of continuous digital workflows. The emergence of standardised web services for geospatial data, sensor network integration and grid technology now offer tools for the creation of such workflows, orchestrated by workflow engines. The creation of continuous digital workflows enables us to create new tools for global collaboration in the earth sciences by integrating the acquisition of data and metadata, and their subsequent processing, modelling and dissemination.

1 Introduction

Intensive research in the earth sciences over the past decades has created a tremendous wealth of literature, data, and sample collections. So far, literature, data and sample collections have been separated and under-utilised due to the absence of continuous digital workflows [1] and the enormous efforts that a digitalisation of analogue data and metadata would cause. Information technology and the internet, in particular web services and grid technology, create the potential for new interpretations for the earth sciences by offering ways to interlink literature, data and samples from the source of data in the field or laboratory, all the way to their interpretation in the literature.

In particular the development of Sensor Web Enablement standards by the Open Geospatial Consortium, in conjunction with grid technology, are now being used to aid environmental monitoring and capture of real-time data in earth observing systems [2]. The Sensor Web Enablement concept extends beyond its application in environmental sensor networks. Its ability to model any data source or process as a sensor that takes in data and puts out processed data. This concept of a "sensor", in conjunction with directory and system management services, makes Sensor Web Enablement a universal tool for earth science data.

In this paper we look at the new tools available to the earth sciences and how they can be used to create continuous digital workflows for monitoring and data capture, data processing and modelling, creation of added value chains, and new forms of publication.

2 Geodata, Web Services and Sensor Networks

Since almost everything in the earth sciences is tied to a geographical location, the use of geospatial data infrastructure suggests itself. Since 1994 the Open Geospatial Consortium (OGC) has created specifications for stan-

dardised web services for geospatial data and several applications have since been developed. These were initially centred on map services but have since expanded into a whole suite of interoperable services.

A recent addition to the OGC geospatial web services is the Sensor Web Enablement (SWE) initiative, a framework of open standards for exploiting Web-connected sensors and sensor systems of all types: flood gauges, air pollution monitors, Webcams, satellite-borne earth imaging devices and countless other sensors and sensor systems [3]. SWE is still under development and several open source applications are already available. A list of currently available SWE components can be found on the OGC website (http://www.opengeospatial.org/).

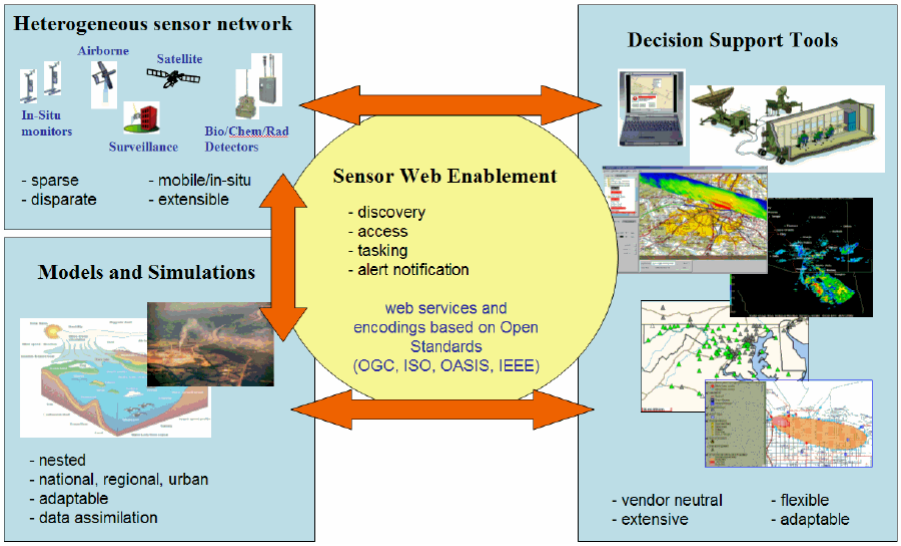

Figure 1: The role of the Sensor Web Enablement framework [3]

SWE describes a sensor as a device that takes in data, processes data and offers the product of its internal processing steps through a standardised interface. In the example of a thermal sensor, the sensor will not measure temperature, but instead it will measure the voltage across a thermocouple, calculate the temperature from the voltage and offer temperature as a numerical data output. A “virtual” sensor does essentially the same: it takes in data, then processes the data and offers a data product through its output interface. The data involved need not to be limited to numerical data, they may also be more complex objects, even entire computational models. This concept of a sensor in SWE makes this technology particularly interesting for the construction of continuous digital workflows in earth science research.

All standard SWE services, except for the Sensor Alert Service (SAS), are servers offering request-response services to client applications. It is not yet possible to subscribe to a regular delivery of data from a SWE service. As a consequence, the output of a SWE Sensor Observation Service (SOS) cannot be used to automatically trigger a next processing step. To chain modular SOS sensors into processing chains therefore requires that the services are orchestrated into a processing chain by a workflow engine. Recent work has shown that workflow engines, based on BEPL [4] or SciFlo [5], can be used for this task.

A particular challenge for the planning and tasking of environmental sensor networks lies in the limited availability of energy and bandwidth. Grid

technology offers services for resource allocation, which could be used to support sensor planning and tasking. The accounting and billing services could then be used for controlling purposes in distributed organisations like virtual observatories.

3 Continuous Digital Workflows

A major challenge facing the use of data in scientific workflow is the absence of continuous digital workflows. There are still numerous breaks between data capture and data processing. Data publication as part of the scientific publication process is still in its early stages [6]. The services introduced in the section above can serve as useful tool both in data acquisition and in data dissemination.

3.1 Web Services for Data Acquisition

Data input into a scientific workflow can occur in a number of ways, but unless data are retrieved from a database, the input data and metadata rarely come without the need of translation from analogue or non-standard digital format into a standardised digital format. The SWE concept bridges this gap. It allows the acquisition of data together with their metadata at their source and thus creates the root of a continuous digital scientific workflow.

The SWE concept of a sensor is very broad and encompasses more than just environmental sensors of remote sensing instruments, it can also model sensors in laboratory instruments or observatories, e.g. mass spectrometers or seismometers. Even data processing algorithms and databases can be modelled as SWE sensors. In this way SWE services can link between “historical” data stored in databases and real-time data collected by sensor networks or observatories. The SWE concept of Observations and Measurements (OM) is ideally suited to integrate the semantically diverse data in earth sciences.

The concepts of an “observation” in OM and the embedded concept of a “feature of interest”, as defined by OGC, are of particular interest to semantically rich disciplines, like the earth sciences. The OGC defines an observation as an act that involves a procedure applied at a specific time and place. The result of an observation is an estimate of some property value. The property itself is associated with the observation domain or feature of interest. The location of the procedure might not be the location of interest for spatial analysis, which makes this model particularly interesting for applications beyond environmental monitoring [7]. The extension of sensor web concepts to the laboratory puts the electronic lab-book into reach as an interoperable module and its integration into continuous digital scientific workflows.

3.2 Continuous Workflows – From Lab to Publication

Scientific scenarios in the earth sciences can be of high semantically richness, since they may span several dimensions, orders of magnitude in time and space, and cross scientific domain boundaries. Here, a study on an underground deposit of methane hydrate in a porous host rock shall serve as a simplified model to illustrate how standard services can be orchestrated by a workflow engine into a continuous digital workflow from data acquisition to publication.

The digital workflow starts with the research for data sources in a catalogue of available data and services and is conducted by the researcher

through a graphical user interface. The catalogue data is provided by an OGC catalogue service (CS-W). This service is not limited to a single source but could aggregate the filtered contents of several other catalogue services (Figure 2).

In our example application, the aim is to model the distribution of solid methane hydrate in the porous host rock. Its behaviour will be influenced by the pressure and temperature conditions in the host rock, and also by the presence of methane or water in the pore space, and other parameters. In this example, one data centre might host the seismic data of the targeted host rocks for methane hydrate deposits, a second data centre might hold a vector or box model of the underground, and a third data centre might hold data on the thermodynamic behaviour of carbon dioxide, which is sequestered in a liquid state. Since marine methane hydrate deposits are characterised by an acoustic transparent zone and an acoustic reflection in parallel to the sea-floor (bottom simulating reflector), seismic profiles could be scanned by image processing services to detect bottom simulating reflectors, and thus identify potential exploration targets.

Once the desired data and services are found, the researcher selects the resources that will be used and edit the workflow in the workflow engine. If applicable, the necessary base data for the numerical model are downloaded through OGC Web Feature Servers (WFS), or other machine-operable download interfaces (Figure 2). Data provided by WFS will be encoded in Geographical Mark-up Language (GML), using the OGC OM concept.

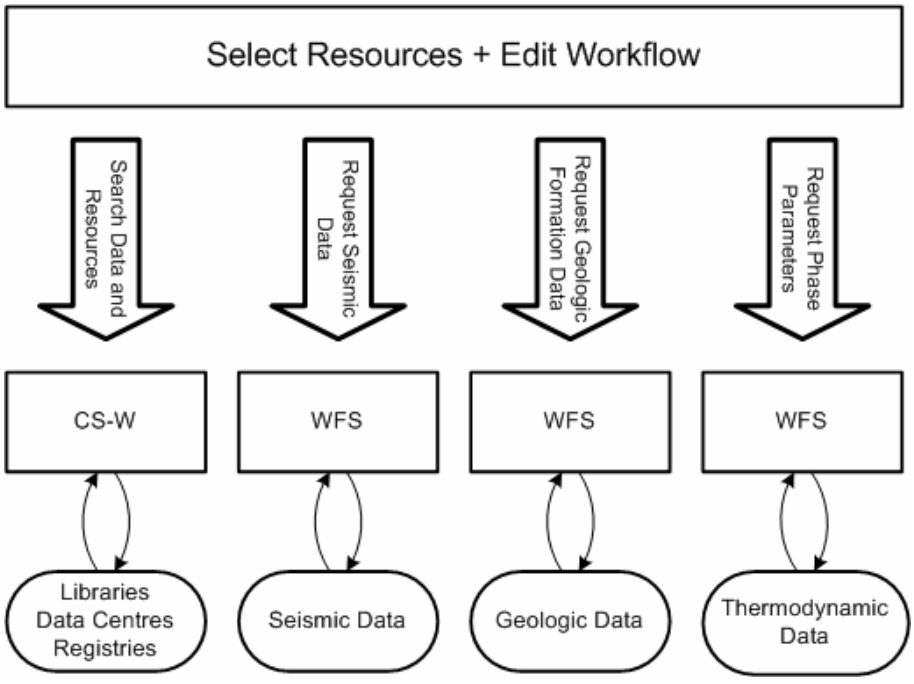

Figure 2: Phase I of a digital workflow. Here, data sources are researched and selected for inclusion into the workflow and linked by processing directives into a workflow chain.

In the second phase (Figure 3), the selected data are used as the base of a model of methane hydrate formation in a porous host rock. In our example, the service that computes the numerical model has been implemented as a complex virtual sensor and fitted with a SWE Sensor Observation Service (SOS) interface. It takes the model base data as “input values” and the result-

ing model as “output values”. These data are, of course, not simple numerical data. The datasets are encoded in GML, making use of the OM data model.

Since numerical models have a high demand for computing resources, this task can be outsourced to a virtualised compute service through a Grid web service. The workflow engine marshals the task into jobs that are sent to the Grid Scheduling Service.

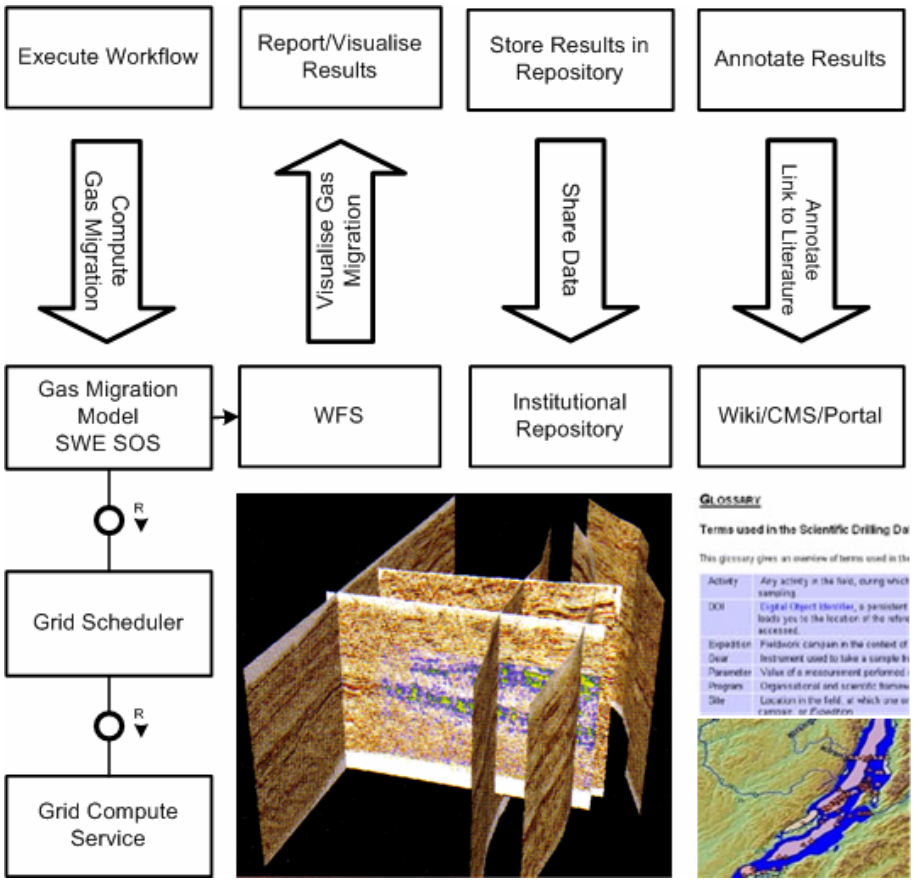

Figure 3: Phase II of a digital scientific workflow. Here, data are processed in a numerical model and the output data provided for reporting and visualisation, storage in an institutional repository. The model results can be annotated and published through a wiki, content management system, or portal.

The results of the numerical simulation are available through a WFS. The result data can be displayed as a report or visualised.

3.3 Web Services for Data Dissemination

Web services have been successfully used for data dissemination. A leading example is the concept of the virtual observatories, which was developed in astrophysics and has now been adopted by other disciplines [8]. However, few of these employ standardised web services. Here, both Open Archive servers and OGC web services offer solutions for interoperable exchange of data and metadata [9]. The Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) is already used by the PANGAEA database to disseminate data to a number of project portals [10]. Client and server applica-

tions for OAI-PMH interfaces are available as Open Source software. In most applications, OAI interfaces are used to disseminate metadata with a Dublin Core profile. Since OAI-PMH does not specify the metadata profile to be used, NASA DIF [11] or ISO 19xxx metadata profiles can be used for dissemination through geological data portals.

Already now most literature is identified by persistent identifiers. It is now possible to assign persistent identifiers to data [6] and soon to geological samples [12]. The use of interoperable web services for data, metadata and scientific literature allow a closer integration of scientific data and scientific literature . Persistent identifiers can be used in publications to reference data in scientific databases or point to the samples, from which the data were derived. Some geological databases already show references from datasets to the literature, in which the data are interpreted and allow references to scientific samples from natural history museums, institutional collections or the core repositories of the international scientific drilling programmes (Figure 4).

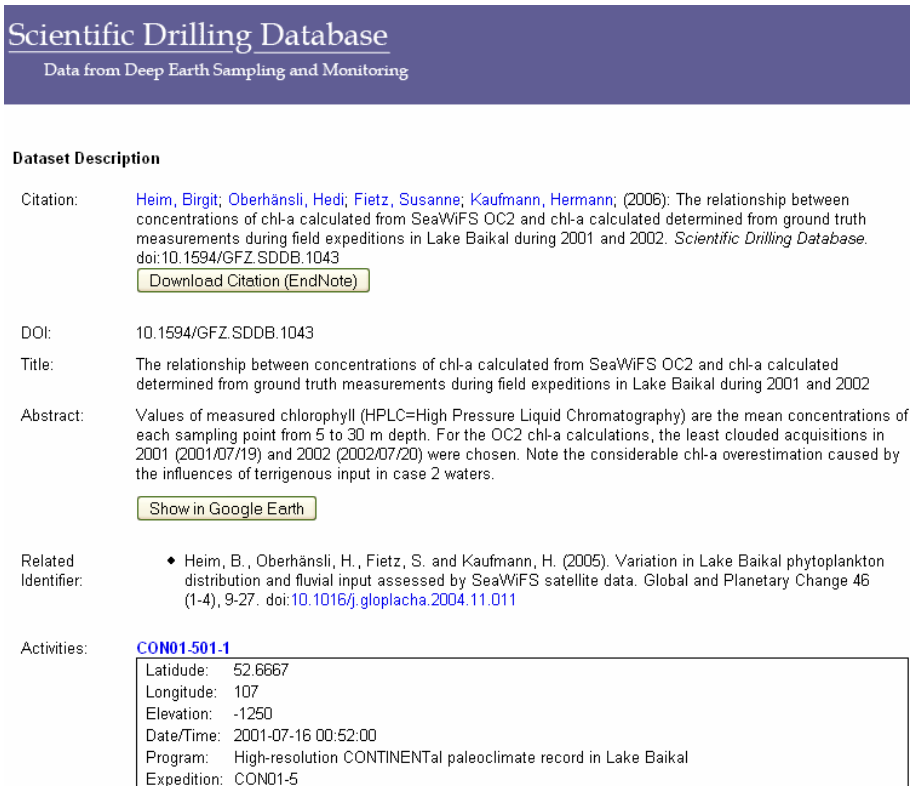

Figure 4: Screenshot of the ICDP Scientific Drilling Database showing the summary of a data publication. The button for data download is off the screen at the bottom of the page. This dataset (doi:10.1594/GFZ.SDDB.1043) points to the journal article of Heim *et al.* (2006) (doi:10.1016/j.gloplacha.2004.11.011), where the data are interpreted.

3.4 Digital Workflows and Grid Technology

Building continuous digital workflows requires that interfaces and services are highly standardised. While access to data and metadata is standardised by OGC services, access to compute and storage services can be stan-

dardised through Grid Web Service interfaces. Especially routine operations that require significant computing power can be outsourced to virtualised Grid resources. Grid services are already used in virtual astronomical observatories [13, 14]. The idea of virtual observatories has already been transferred to the earth sciences [15]

While Open Access to knowledge and data in the sciences and humanities has become a widely debated topic, work in progress must be protected to secure the intellectual property of the researcher. Also, maintaining infrastructure for scientific research comes at a cost. Therefore it is important that access to resources can be restricted, if necessary. Grid technology offers services to enforce access policies, as well as for controlling and billing their use. However, the social dimension, how this will impact on the management of virtual organisations, is yet to be determined.

The transition to digital scientific workflow solves many problems in our present way of handling scientific data and literature, but it also poses new challenges. One of the greatest challenges is the question of long-term preservation, which is both technological and organisational questions. While interoperable web services and grid technology provide tools that can provide a technological solution, data stewardship requires changes in the scientific process.

4 Conclusions

A continuous digital scientific workflow is a crucial tool to enable global collaboration in the earth sciences. Today, many data are lost or not reusable due to incompatible media in the scientific workflow and loss of descriptive metadata. The Open Geospatial Consortium reference architecture provides standardised services and standard formats for data and metadata that can be used as elements of a continuous digital workflow. Particularly useful are Web services based on the emerging concept of OGC Sensor Web Enablement, grid technology and interoperable institutional repositories for scientific literature. These elements can orchestrated into continuous digital workflows for modelling and scholarly communication. The creation of continuous digital workflows from the laboratory, through analysis and discussion, into the scholarly communication in the scientific literature will greatly improve the utilisation of existing data and literature in the earth sciences and thus enable new discoveries.

References

1. Oldow, J.S., et al., *Digital Acquisition, Analysis, and Visualization in the Earth Sciences.* EOS, Transactions, American Geophysical Union, 2006. **87**(35): p. 351.
2. Pierce, M. *Integrating Geographical Information Systems and Grid Applications*. in *Sixth Annual NASA Earth Science Technology Conference - ESTC2006*. (2006).
3. Botts, M., et al., *OGC Sensor Web Enablement: Overview And High Level Architecture*, in *OpenGIS White Paper*. Open Geospatial Consortium, Inc.: Wayland, MA. 2006.
4. Emmerich, W., et al., *Grid Service Orchestration Using the Business Process Execution Language (BPEL).* Journal of Grid Computing, 2005. **3**(3): p. 283-304.
5. Wilson, B.D., et al., *Collaborative Science Using Web Services and the SciFlo Grid Dataflow Engine.* EOS, Transactions, American Geophysical Union, 2006. **87**(52, Fall Meet. Suppl.): Abstract IN42A-03.

6. Klump, J., et al., *Data publication in the Open Access Initiative.* Data Science Journal, 2006. **5**: p. 79-83.
7. Cox, S., *Exchanging observations and measurements: Applications of a generic model and encoding.* EOS, Transactions, American Geophysical Union, 2006. **87**(52, Fall Meeting Suppl.).
8. Golombek, D., *Archives, databases and the emerging virtual observatories.* Astronomy & Geophysics, 2004. **290**: p. 449-456.
9. Bekaert, J. and H. Van de Sompel, *Augmenting interoperability across scholarly repositories*. 2006, Andrew W. Mellon Foundation: New York, NY. p. 17.
10. Schindler, U., B. Bräuer, and M. Diepenbroek. *Data Information Service based on Open Archives Initiative Protocols and Apache Lucene*. *GES 2007*. Baden-Baden, Germany. 2007.
11. *Directory Interchange Format (DIF) Writer's Guide*, in *Global Change Master Directory*. 2006, National Aeronautics and Space Administration.
12. Lehnert, K., et al., *The Digital Sample: Metadata, Unique Identification, and Links to Data and Publications.* EOS, Transactions, American Geophysical Union, 2006. **87**(52, Fall Meet. Suppl.): Abstract IN53C-07.
13. Walton, N.A. and E. Gonzales-Solarez, *AstroGrid: A place for your science.* Astronomy & Geophysics, 2006. **47**(6): p. 3.22-3.24.
14. Williams, R., *Grids and the virtual observatory*, in *Grid Computing - Making the Global Infrastructure Reality*, F. Berman, T. Hey, and G.C. Fox, Editors. Wiley & Sons, Ltd.: New York, NY. 2003, p. 409-435.
15. Fusco, L. and J. van Bemmelen, *Earth observation archives in digital library and grid infrastructures.* Data Science Journal, 2004. **3**: p. 222-226.