Adv. Geosci., 8, 3–9, 2006
www.adv-geosci.net/8/3/2006/

Advances in Geosciences

Mediation to deal with information heterogeneity – application to Earth System Science

L. Bigagli[1,2], **S. Nativi**[1,2], **and P. Mazzetti**[1,2]

[1]Institute of Methodologies for Environmental Analysis of the National Research Council (IMAA-CNR), Area della Ricerca di Potenza, Contrada Santa Loja, Zona Industriale, I-85050 Tito Scalo (PZ), Italy
[2]University of Florence, Piazza dell'Università, I-59100 Prato (PO), Italy

Received: 15 September 2005 – Revised: 18 January 2006 – Accepted: 10 March 2006 – Published: 6 June 2006

Abstract. We address the problem of data and information interoperability in the Earth System Science information domain. We believe that well-established architectures and standard technologies are now available to implement data interoperability. In particular, we elaborate on the mediated approach, and present several technological aspects of our implementation of a Mediator-based Information System for Earth System Science Data. We highlight some limitations of current standard-based solutions and introduce possible future improvements.

1 Introduction

Heterogeneity occurs by nature, in any given universe of discourse, under many aspects, e.g. technical, geographical, temporal, administrative, linguistic, cultural, social (see Wiederhold, 1993, and Busse et al., 1999, for a classification of heterogeneity facets).

For example, heterogeneity issues are remarkable in education, which implies the exchange of information between subjects that are usually quite distant from each other, under several points of view (e.g. their level of instruction).

Another such arena is scientific research, entailing collaborative efforts and information exchange between members of highly heterogeneous (although hopefully "convergent") communities. Particularly, Science Digital Libraries must support interdisciplinary exchange of information both in research and in educational settings (Bartolo et al., 2004).

In the Earth System Science domain, the vastness of the topic and the diversity of applications have led to a tremendous heterogeneity of resources and procedures, which hinders cooperation among the different actors and stakeholders.

Correspondence to: L. Bigagli
(bigagli@imaa.cnr.it)

There is an impressive record of efforts to oppose heterogeneity by homogenizing the universe of discourse, which is normally unfeasible, or undesirable, or impossible.

Heterogeneity issues are more successfully mitigated by the implementation of interoperability solutions.

In the following sections, we relate the above considerations to Information and Communication Technology (ICT) concepts and literature, such as the theory of Mediator-based Information Systems, focusing on its application to the Earth System Science information domain.

1.1 Data and information interoperability

Interoperability may be generally defined as *the ability to access, exchange, relate and combine information from multiple heterogeneous sources* (see Wiederhold, 1999). Information may be defined as *knowledge acquired through study or experience or instruction*. In the usual connotation of language as a persistent representation of knowledge, interoperability issues basically become *language* issues, and may be conveniently addressed at the three levels of lexicon, syntax and semantics.

Knowledge, in turn, may be defined as *the psychological result of perception and learning and reasoning*. According to these definitions, information is always a subjective concept, whose meaning depends on the context and whose value is apparent to the receiver (who/which is actually part of the context itself).

If the receiver is a human, a problem that arises is our lack of scalability in managing increasing volumes of information. This is referred to as *information overloading* (see Miller, 1956, on human cognitive limitations).

This problem is most evident when we face machines, for instance when we have to digest the overwhelming results of a web search. It is also a very current issue in education, where the learning capabilities of students, as well as

Published by Copernicus GmbH on behalf of the European Geosciences Union.

the effectiveness of our teaching methods, must cope with the exponential growth of networked sources of knowledge.

According to Mitra et al. (2005), techniques to mitigate information overloading (e.g. providing pertinent results to queries, or ranking results better, or mining data more accurately) will be the real drivers of Internet evolution. In fact, research about search engines is probably the roughest battleground of the Internet industrial war.

Different representations may be derived from given information, for example to optimize transmission or storage. A representation of information in a form suitable to be automatically managed is referred to as *data*. The process used to obtain such form is referred to as *encoding*.

Data-processing systems provide the capability to restore (*decode*) the original representation of information, which may be considered as the "meaning" of that data. Likewise, a data-processing system may be considered as the context where that data is meaningful.

Thus, it is possible to distinguish between data and information interoperability. In the above terminology, they refer to the capacity to move data and information across heterogeneous sources and destinations, in such a way that the receiver is able to restore the original representation of information (i.e. data) and understand its meaning, respectively.

Architectural variants are possible, for example, in the cardinality of senders and receivers (distribution/integration of information), in the communication paradigm (client-server, peer-to-peer, etc.), or in the messaging pattern (request-response, one-way, publish-subscribe, etc.).

In the end, it is possible to say that a significant difference between information and data resides in the interpreting context and resembles the difference between humans and machines: basically, just a richer context. The more is our context formalized and adapted to automatic management, the greater will be our ability to overcome information overloading and to bridge the unavoidable heterogeneities of the real world.

1.2 Mediator-based information systems

A mediator-based information system is a federation of disparate (i.e. heterogeneous and distributed) systems that relinquish some of their autonomy for the sake of collaboration, and interoperate by abstracting their own local data model to a common one, the *federal model*. Each federation member may implement the common model just as a virtual view over the existing internal one. Thus, members can seamlessly participate in multiple federations, with different data models, requirements and features, by providing appropriate, conforming interfaces: for example, a storage system can simultaneously participate in both a Unix and a Windows distributed file system (e.g. by implementing respectively the Network File System and the Server Message Block protocol for sharing files, printers, serial ports, etc.)

Members must strictly conform to the federal model when communicating with each other, but can otherwise maintain their internal structure and a large degree of autonomy. With this regards, mediator-based federations differ from loose-federations, that have no explicit common policy. They also differ from distributed database solutions, that mandate a uniform internal structure (Busse et al., 1999). Sometimes, successful federal models may be adopted by federation members also as their own internal models. This avoids the need for mediation and imposes a unique model which overcomes heterogeneity: an example is the spreading of Intranets solutions. They are based on TCP/IP application protocols which, in the past, were used instead to enable interoperability between heterogeneous proprietary solutions.

A federal model is usually an instance of the hierarchical, semi-structured data model, which can be easily mediated to traditional relational and object-oriented models, as well as to file systems and generally to web data, which are intrinsically semi-structured (Busse et al., 1999).

Since a federal model may be more general than a local one, loss of information may occur in the abstraction process. Hence, mediator-based information systems are targeted at read-only applications. On the other hand, federated models tend to be very detailed, to encompass different approaches avoiding oversimplification. Thus, information overloading is a common issue for federation members.

Besides, the absence of requirements on the internal structure implies restrictions to query capabilities, which are usually limited to a set of pre-canned queries.

More generally, the mediated approach relies on the identification of *articulation points* around a particular heterogeneity boundary and the implementation of adaptation logic, whose execution is delegated to a specialized, lightweight component: the Mediator (Wiederhold, 1992). For example, proprietary e-mail services, differing for message format and network transport protocols, define heterogeneous e-mail domains. Specific systems, namely e-mail gateways, located at the domain boundaries, are in charge of mediating messaging traffic, making e-mail systems interoperable.

According to the mediated approach, data model integration (which is an aspect of data interoperability) can be easily achieved by adapting the source data model to the destination one, before data exchange.

In ICT, this solution has been conceptualized as a *structural* design pattern, that is a way of arranging software components to solve common issues: the well-known Adapter pattern (Gamma et al., 1995).

As shown in Fig. 1, the Adapter wraps a component (the Adaptee), exposing to a Client a Target interface, with which it is capable of interacting. The Adapter then mediates between the Client and the Adaptee, transforming the Client requests in a suitable form for the Adaptee to respond.

A common real-life example of this pattern is language interpretation between interlocutors in a political meeting: each politician is assisted by interpreters (the Adapters), one

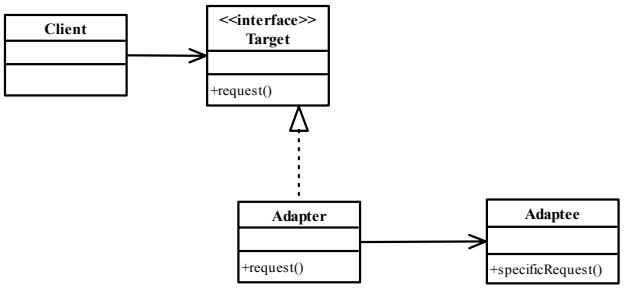

Fig. 1. Adapter pattern class diagram.

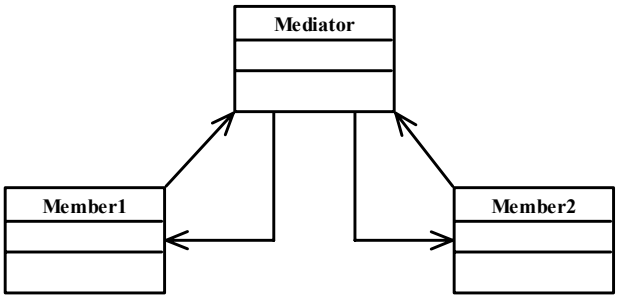

Fig. 2. Mediator pattern class diagram.

for each foreign language in use, translating questions and answers to his/her native language.

In general, given n different systems to interconnect, n^2-n different adapters are necessary.

Mediator-based federations avoid the proliferation of adapter components, defining a particular, possibly virtual (i.e. not used internally) federal model as a common ground where all participants interactions are materialized.

In ICT, this approach has been conceptualized as a *behavioural* design pattern, that is a collaborative arrangement of software components to perform a task that no single component can carry out alone: the Mediator pattern (Gamma et al., 1995).

As shown in Fig. 2, any communication between the federation Members is routed through the Mediator component, which is able to apply the appropriate common policies.

Adaptation may still be necessary for those participants whose local model differs from the federal one. However, given n different system to be interconnected, at most n different adapters are sufficient.

According to the partitioning introduced above, lexical and syntactic adaptation is usually demanded to a specialized subcomponent, called Wrapper. The Mediator performs higher-level, semantics adaptations.

The implementation of a Mediator-based federation may result from a bottom-up process, by the spontaneous adoption of a system-wide policy, that becomes a *de facto* standard when a critical mass is reached. On the other hand, ,a formal standardization is also possible, implementing a top-down process. This usually requires extensive resources in planning, analysis and negotiation to identify and conceive the federal model.

In keeping with the previous linguistic example, the top-down approach was attempted by Doktoro Esperanto (Dr. Hopeful) in the frustrated effort to establish a universal and artificial language (Zamenhof, 1887). An example of the bottom-up approach is the current spreading of the de facto standard English (more precisely the Globish: Global + English) language.

1.3 Mediation applied

Mediation has been successfully applied to the task of interconnecting diverse communication systems and enabling seamless data transport. Internet standards and specifications support mediating solutions for extending local communication services to a worldwide network. Architectural solutions based on intermediate systems (i.e. gateways) can be adopted to federate different communication systems (e.g. messaging systems such as e-mail, SMS, etc.) using Internet specifications. In particular, the Internet global addressing schema (namely IP addressing), RFC 822 and MIME message formats, are examples of such enabling specifications.

Internet specifications have been a very successful federal model, to the point that, nowadays, most local communication systems are based on them. In particular, the Web architecture, based on URI (Uniform Resource Identifier) identification schema, standard application level protocols (mostly HTTP which is defined in RFC 2616 as “a generic, stateless, protocol which can be used for many tasks beyond its use for hypertext, such as name servers and distributed object management systems”) and specifications (e.g. URL-encoding for parameter passing) is widely adopted as a federal model for interoperability of different application domains.

Public Key Infrastructures (PKIs) are another example of mediation-based federations: instead of entrusting one another, community members rely on an intermediary (the Certification Authority) to adapt mutual trust levels and implement collective policies, such as Single-Sign-On.

Recently, XML (Bray et al., 2001) has emerged as a federal model for lexical and syntactic data interoperability. The impact of XML has been crucial, because it is a natural encoding for the semi-structured data model. Moreover, the powerful design of XML has enabled the definition of a growing number of specialized dialects to support the most diverse tasks, including the classical aspects of data management: Data Definition Language (DDL) as the XML Schema Language; Data Query Language (DQL) as XQuery and XPath; Data Manipulation Language (DML) as the eXstensible Stylesheet Language Transformations (XSLT) and XUpdate (Clark, 1999).

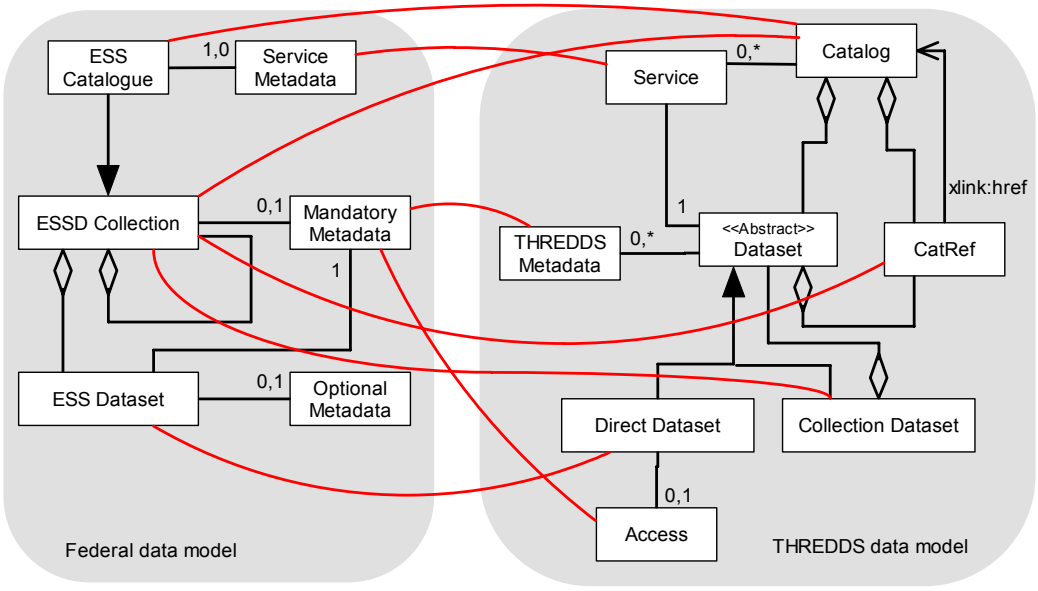

Fig. 3. THREDDS model mapping to the federal model.

Furthermore, XML-related technologies such as Web Services and the so called XML Protocol (SOAP, WSDL, etc.) provide an interoperable representation of procedural constructs such as function calls, transactions, operators, etc. Noticeably, XML adoption in data-management systems backend is also gaining momentum, with Native XML Databases (NXDs).

Another successful application of the mediating approach is Java technology, built around the specification of an abstract, virtual platform (the Java Virtual Machine) and its associated machine language (the Java bytecode), adapting the actual services of heterogeneous operating systems and hardware.

Together, Java and XML, the "attractive" technologies of a few years ago (Wiederhold, 1999), are now mature solutions to the problem of data interoperability, tackling lexicon and syntax mismatches and balancing computational load among data providers and receivers over the global Internet. However, these technologies are quite limited in expressing semantics: when dealing with *information* interoperability, higher-level mediation tools and technologies are needed.

This is particularly evident in complex contexts, where agreement on standard technological solutions is difficult to achieve. One such context is discussed in the next section: the Earth System Science.

1.4 Application to Earth System Science

As far as data model and interoperability interfaces are concerned, the Earth System Science community has achieved a certain maturity. Currently, the momentum of Geomatics standardization initiatives, such as those of OGC and ISO TC 211, is boosting the implementation of global interoperability solutions for Earth System Science, like GEOSS (Global Earth Observation System of Systems) [1] and the European GMES (Global Monitoring for Environment and Security)[2] and INSPIRE (INfrastructure for SPatial InfoRmation in Europe) initiatives[3].

Several XML dialects are being developed by the diverse Information Communities to exchange data between the community members, e.g. ESML, ncML (Nativi et al., 2005b), GML. A growing number of mediation tools and services supports the conversion between different markup languages, enables access to data by a variety of users and provides a framework for further extensions (see Bartolo et al., 2004, about the Science Digital Libraries domain). In addition, several initiatives try to promote interoperability among heterogeneous communities, such as the GALEON network[4], which aims to bridge the conceptual models of GIS and netCDF communities (Domenico et al., 2006). It is noteworthy the experimentation of a specific mediation language, ncML-GML (Nativi et al., 2005a), to achieve the models reconciliation.

In the past years, we have successfully experimented with a Mediator-based federation of disparate sources of Earth System Science Data (ESSD), featuring technologies such as HTTPS for security mediation, Java for execution environment mediation, and XML for data model mediation (Nativi et al., 2001; Nativi et al., 2003). More recently, we have

[1] http://earthobservations.org/
[2] http://www.gmes.info/
[3] http://inspire.jrc.it/
[4] http://www.opengeospatial.org/initiatives/?iid=173

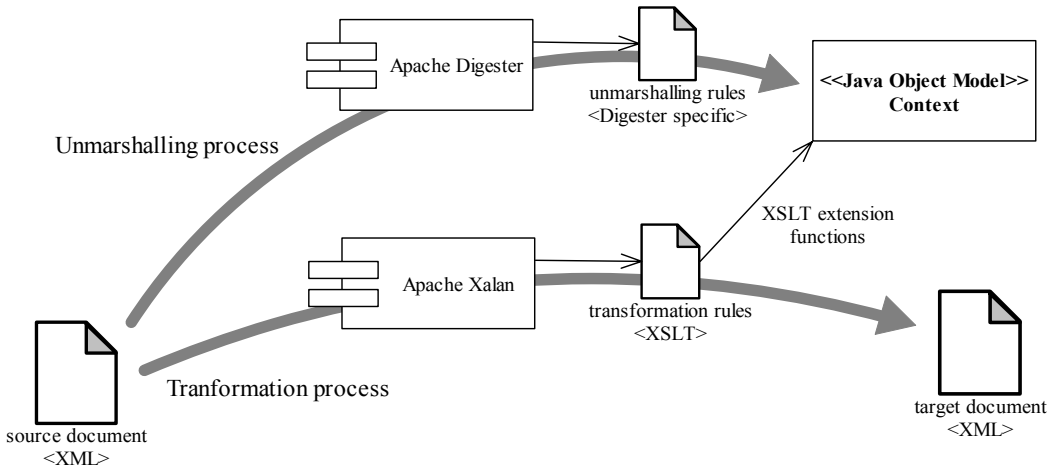

Fig. 4. Technologies involved in the transformation process.

implemented this solution as a Service-Oriented framework of modular components, taking part in a SDI experiment (Nativi et al., 2004; Bigagli et al., 2004).

The proposed solution includes discovery services for ESSD, relating and combining data from sources characterized by heterogeneous models and encodings, and presenting the users a uniform federal view of metadata and data, based on the ISO19100 standard series. Supported participant data models include Unidata THREDDS[5], ESA EOLI[6], OGC WCS (OpenGIS, 2006), and others. Figure 3 illustrates the main relationships among the federal data model (on the left) and a participant local model: the THREDDS data model (on the right).

A mediator-wrapper component transforms the flowing XML resources from the local to the federal schema, by means of ad-hoc structural mappings. We have implemented the transformation process using a mixture of declarative and procedural constructs expressed in XSLT and in Java, respectively. Declarative constructs allow to transform an XML resource by describing (declaring) the result in terms of its expected structural properties, without having to specify the transformation algorithm. As the name suggests, XSLT modifies the form of the involved resources, not their substance. Thus, it supports transcoding, reordering, deleting and other such lexical and syntactical rearrangements.

Procedural constructs support more complex mappings, such as those implied by algorithmic decisions, impossible or unpractical to express in XSLT. Figure 4 depicts the transformation process, with regards to the technologies involved.

Expectedly, we had to consider the issue of information overloading when we designed the federal model for the metadata of a generic geospatial information resource. The model is based on the widely accepted ISO 19115 standard (ISO, 2003). ISO comprehensive metadata set includes 300 entries, and it is very unlikely that so much ancillary information ever be needed by the average data producer and consumer. Actually, according to our experience, even some mandatory metadata of the ISO 19115 core set may be considered superfluous, for example as far as data discovery is concerned. In addition, highly specialized information communities may have difficulties to deal with other mandatory metadata. On the other hand, these communities often need domain-specific information not included in the general-purpose standard.

Ideally, to improve usability and alleviate (ancillary) information overloading on the user side, client applications should adapt to every single user preferences about metadata element profiling (e.g. element ranking, alias, nesting level, etc.) With this regards, we have designed and implemented a "metadata profiling" feature, by means of a meta-XSLT technique, used to generate a customized XSLT, used in turn to select and render the metadata subset considered relevant by the user, possibly customizing their order, nesting level, alias and graphical appearance.

To address these issues in the general case, ISO 19115 specifies standard rules for metadata profiling, pruning unneeded information as well as extending it appropriately.

As shown in Fig. 5, the envisioned use-case is for a given user community to select a subset in between the core and the comprehensive metadata set, complementing it with domain-specific extensions.

Although the definition of user community profiles is a suitable strategy to improve data completeness (by adding all relevant information) and accuracy (by removing irrelevant information), it is questionable whether it could ultimately undermine data interoperability.

ISO 19115 is apparently anticipating the problem, with the elementary recommendation that "prior to the creation of extended metadata a careful review of the existing metadata

[5]http://www.unidata.ucar.edu/projects/THREDDS/
[6]http://odisseo.esrin.esa.it/eoli/

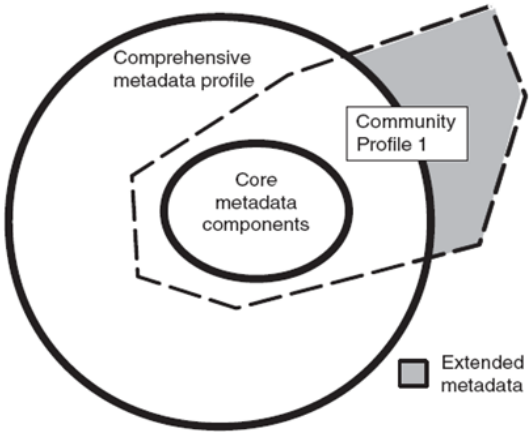

Fig. 5. ISO 19115 profiles (source: ISO, 2003).

within this International Standard must be performed to confirm that suitable metadata does not already exist" (ISO, 2003). According to Mitra et al. (2005), "in the end, we will have to live with inconsistencies" and "Negotiation is futile". Even not going this far, it is arguable that the growing adoption of ISO 19115 will imply the parallel diffusion of a plethora of profiles, promoted by the different stakeholders, that will eventually face a smaller-scale version of just the same interoperability issues they have always been confronting.

Given the substantial uniformity of the data model, most mismatches may then be solvable at a lexical and syntactic level, e.g. with a more or less sophisticated application of XSLT.

However, our experience indicates that the very frequent modifications of the existing technical baseline (including specifications and implementations alike) could jeopardize a strategy based on syntactic and lexical mapping, and pose severe maintainability problems. This is particularly true considering the scarce availability of tools to ease working with XSLT and the present restrictions of the declarative approach.

The optimal solution would be to work at a higher level of abstraction: the conceptual level, which is nowadays fairly consolidated and captured in well accepted standards. This topic is highly investigated and the most promising solutions entail the definition of XML ontologies, by means of the Web Ontology Language (OWL), and their automated mapping through customized methods or algebraic operations (Bermudez, 2004; Mitra et al., 2005). It is envisioned that context formalization could ultimately enable the automatic management of "meaningful" data across the World Wide Web. In this scenario, usually referred to as the Semantic Web (Berners-Lee et al., 2001), mediators based on semantics-aware technologies would allow to map and crosswalk among concepts, disregarding the details of encoding and thus achieving a real information integration between different information communities (Pazienza et al., 2005).

2 Conclusions and future directions

We have addressed the problem of data and information interoperability in the Earth System Science information domain.

In particular, data interoperability is the capacity to move data and information across heterogeneous sources and destinations, in such a way that the receiver is able to restore the original representation of information. This problem can be tackled through the mediated approach, that is identifying the existing heterogeneity boundaries and implementing suitable adaptation logic by means of specialized, lightweight components.

We have experimented with a mediation-based federated solution based on Java and XML, which has proven suitable for implementing data interoperability in the Earth System Science domain. The proposed federal model is based on the ISO TC 211 conceptual model, which is nowadays fairly consolidated and captured in well accepted standards.

However, our experience indicates that the complexity and variability of the existing technical baseline could pose severe maintainability problems, which reflect the intrinsic heterogeneities of the disparate, autonomous systems that produce and manage geospatial information, ultimately related to the Earth's worldwide extension.

Our capacity to bridge the unavoidable heterogeneities of the real world could be improved if a larger part of our context was formalized and adapted to automatic management. In fact, the application of semantics-aware technologies seems promising to track the evolution of standard solution and ease maintainability and evolution. The envisioned implementation of the Semantic Web seems to be the only solid starting point for real information interoperability in the Earth System Science domain.

Edited by: E. Cutrim, M. Ramamurthy, S. Nativi, and L. Miller
Reviewed by: anonymous referees

References

Bartolo, L. M., Cole, T. W., Giersch, S., Wright, M.: NSF/NSDL Workshop: Scientific Markup Languages Report, National Science Foundation, Arlington, Virginia, June 14–15, 2004.

Bermudez, L. E.: ONTOMET: Ontology metadata framework, Ph.D. thesis, Drexel University, Philadelphia, USA, December 2004.

Berners-Lee, T., Hendler, J., and Lassila, O.: The Semantic Web, Scientific American, May 2001.

Bigagli, L., Nativi, S., Mazzetti, P., and Villoresi, G.: GI-Cat: a Web Service for Dataset Cataloguing Based on ISO 19115, Proc. of 1st International Workshop on Geographic Information Management (GIM'04) – 15th International Workshop on Database and Expert Systems Applications (DEXA'04), IEEE Computer Society Press, ISBN 0-7695-2195-9, Saragozza, Spain, 30 August–3 September 2004, pp. 846–850, 2004.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F.: Extensible Markup Language (XML) 1.1. W3C Recommendation, February 4, 2004.

Busse, S., Kutsche, R.-D., Leser, U., and Weber, H.: Federated Information Systems: concepts, terminology and architectures, Technical Report Nr. 99-9, TU Berlin, 1999.

Clark, J.: XSL Transformations (XSLT) Version 1.0. W3C Recommendation, November 16, 1999.

Domenico, B., Nativi, S., Caron, J., Bigagli, L., and Davis, E. R.: A standards-based, web services gateway to netCDF datasets, Proc. of AMS – 22nd IIPS Conference, Atlanta, Georgia, abstr. no. 8.1, February 2006.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J.: Design Patterns, 1st ed., Addison-Wesley Professional, January 15, 1995.

ISO/IEC 19115:2003, Geographic information – Metadata, ISO International Standard, Geneva 2003.

Miller, G.: The Magical Number Seven±Two, Psych. Rev., 68, 81–97, 1956.

Mitra, P., Wiederhold, G., and Decker, S.: Dealing with Semantic Interoperation of Data, IEEE-CS International Symposium Global Data Interoperability – Challenges and Technologies, Sardinia, Italy, June 20–24, 134–144, 2005.

Nativi, S., Mazzetti, P., Bigagli, L., and Giuli, D.: Interoperability Federated System for the Scientific Community working in the EOS Sector, Proc. IEEE 2001 International Geoscience and Remote Sensing Symposium – IGARSS '01, ISBN 0-7803-7031-7, Sydney, Australia, 9–13 July 2001, Vol. 3, pp. 1185-1187, 2001.

Nativi, S., Mazzetti, P., Bigagli, L., and Giuli, D.: Federating EOS Heterogeneous and Distributed Information Resources, Atti AMS – 19th IIPS Conference, Long Beach, California, abstr. no. 1.48, February 2003.

Nativi, S., Bigagli, L., Mazzetti, P., and Cuomo, V.: Applying SOA to Earth Observation: the COS(OT) experience, Proc. of ICTTA '04, IEEE Press, ISBN 0-7803-8482-2, Damascus, Syria, 19–23 April 2004, pp. 323–324, 2004.

Nativi, S., Caron, J., Davis, E., and Domenico, B.: Design and implementation of netCDF Markup Language (NcML) and its GML-based extension (NcML-GML), Computers and Geosciences, November 2005a.

Nativi, S., Domenico, B., Caron, J., Davis, E., and Bigagli, L.: An interoperability language to connect netCDF and Geographic communities: ncML-GML v. 0.3.2. GML and Geo-Spatial Web Services Conference 2005, Vancouver, Canada, July 18–22, 2005b.

OpenGIS Consortium, Inc.™: "Web Coverage Service (WCS), Version 1.0.0", OGC 03-065r6, 2006.

Pazienza, M. T., Stellato, A., Henriksen, L., Paggio, P., and Zanzotto, F. M.: Ontology Mapping to support ontology-based question answering, Proc. 2nd MEANING Workshop Trento, Italy, February 2005.

Wiederhold, G.: Mediators in the Architecture of Future Information Systems, IEEE Computer, pp. 38–49, March 1992.

Wiederhold, G.: Intelligent Integration of Information, ACM-SIGMOD 93, Washington D.C., pp. 434–437, May 1993.

Wiederhold, G.: Mediation to Deal with Heterogeneous Data Sources. Interoperating Geographic Information Systems, Springer LNCS 1580, Proc. Interop'99, Zurich, pp. 1-16, 1999.

Zamenhof, L. L. (alias Doktoro Esperanto): Unua Libro, Warsaw, July 26, 1887.