Adv. Geosci., 8, 83–90, 2006
www.adv-geosci.net/8/83/2006/

Advances in Geosciences

Data integration with the Climate Science Modelling Language

A. Woolf[1], B. Lawrence[2], R. Lowry[3], K. Kleese van Dam[1], R. Cramer[3], M. Gutierrez[2], S. Kondapalli[3], S. Latham[2], D. Lowe[2], K. O'Neill[1], and A. Stephens[2]

[1]CCLRC e-Science Centre, UK
[2]British Atmospheric Data Centre, UK
[3]British Oceanographic Data Centre, UK

Received: 30 October 2005 – Revised: 9 January 2006 – Accepted: 25 January 2006 – Published: 6 June 2006

Abstract. The Climate Science Modelling Language (CSML) has been developed by the NERC DataGrid (NDG) project as a standards-based data model and XML markup for describing and constructing climate science datasets. It uses conceptual models from emerging standards in GIS to define a number of feature types, and adopts schemas of the Geography Markup Language (GML) where possible for encoding.

A prototype deployment of CSML is being trialled across the curated archives of the British Atmospheric and Oceanographic Data Centres. These data include a wide range of data types – both observational and model – and heterogeneous file-based storage systems.

CSML provides a semantic abstraction layer for data files, and is exposed through higher level data delivery services. In NDG these will include file instantiation services (for formats of choice) and the web services of the Open Geospatial Consortium (OGC).

1 Introduction

NERC DataGrid (NDG) (Lawrence et al., 2003; Woolf et al., 2005a) is a UK-funded project to develop a Grid-based infrastructure for uniform discovery of, and access to, a wide range of environmental data across the UK and beyond. It will provide a large-scale integration platform for widely distributed and heterogeneous data sources. Focussing initially on the curated archives of the British Atmospheric Data Centre (BADC) and British Oceanographic Data Centre (BODC), eventual deployment is planned across other of the "designated data centres" of the UK's Natural Environment Research Council (NERC). In addition, peering (both for data discovery and access) with international sister projects is underway.

Data heterogeneity in NDG relates to both the data type and storage format. Data types include, for instance, single point observations (e.g. a rainfall measurement), collections of measurements from coordinated field campaigns (e.g. a marine science cruise) and operational instruments (e.g. weather radars), and *tera*-scale numerical simulations from state-of-the-art climate models. Apart from the heterogeneity of data type, the storage format of data in NDG varies tremendously. Both relational and file-based data sources are present, and numerous file formats are used – even for the same data types. A sophisticated integration framework is needed to present a uniform interface to such a diverse range of data. The overall architecture and functionality of NERC DataGrid has been described by Lawrence et al. (2003, 2004); O'Neill et al. (2003); Woolf et al. (2004).

Some key requirements for data integration in NDG include:

- **"In-place" integration:** The system must leave data in-situ and utilise existing storage systems and formats.
- **Scalability:** Deployment effort should scale with new data providers, and significant re-engineering should be avoided for new storage mechanisms or delivery interfaces.
- **Enhancing access:** Greater return on investment and increased collaboration may be realised by supporting non-traditional interfaces such as GIS.

From the above considerations, an information integration architecture emerges based on a common intermediate data model providing an abstraction layer between legacy storage structures and exposed interfaces (including standards-based delivery modes such as those of the Open Geospatial Consortium). Fundamentally, information communities are defined

Correspondence to: A. Woolf
(A.Woolf@rl.ac.uk)

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

by shared conceptualisations (or "semantics"), and so semantic models of data are used in NDG to provide the integration key. A common semantic data model provides a uniform language across providers and users, avoiding details of storage artefacts or client software requirements. It supports a wrapper/mediator architecture so that important, sufficiently standard, software interfaces may be provided without requiring significant effort on the part of providers.

2 Geospatial data standards

An emerging series of international standards for geospatial data, metadata and services provides a timely basis for meeting the data integration requirements outlined above (Woolf et al., 2005b). The series (comprising some 40 standards) is being developed by Technical Committee 211 (Geographic Information and Geomatics) of the International Organisation for Standardisation (ISO)[1]. We introduce below key elements of the framework.

2.1 Domain Reference Model

The overall scope of the series of standards is outlined in ISO 19101 (2002). This standard introduces the "Domain Reference Model". This is an abstract information architecture for geospatial data infrastructres, and provides the basis for standardisation addressed through other standards in the series.

At the core of the model is a geospatial "Dataset". A Dataset contains "Feature" instances (see Sect. 2.2 below) and related objects, and is described by "Metadata". "Geographic information services" operate on a Dataset, while the logical structure and semantic content of a Dataset is described through an "Application schema" (see Sect. 2.3 below).

2.2 "Features"

A geographic "feature" is defined by the ISO TC211 standards as an "abstraction of real world phenomena". Any important entity from a universe of discourse may be characterised as a "feature" in terms of its attributes, associations with other features, and operations that may be performed (the so-called "General Feature Model", ISO 19109 (2005)). In essence, features provide an object view of data[2], and may occur as both types and instances. Feature type definitions may be stored for re-use in catalogues (ISO 19110, 2005). Since features encapsulate important data semantics within communities of practice, such Feature Type Catalogues may be regarded as 'semantics repositories' within an overall information architecture. Feature instances are the primary constituents of geographic Datasets.

2.3 Application schema

While a Dataset logically comprises a collection of feature instances, its precise structure is described through an "Application schema" (ISO 19101, 2002; ISO 19109, 2005). The Application Schema specifies the allowable feature types (and their cardinalities) that may be contained, and any relationships between them. It also specifies what metadata and data quality information (if any) should be contained within a Dataset. An Application Schema may define its own Feature Types, or else refer to external Feature Type Catalogues.

2.4 Governance issues

The integration framework outlined above aims to capture semantics of important community information types. To be successful, any attempt to apply this framework must engage with the community in question. Various points of agreement must be established: (1) what are the information objects that should be modelled as features? (2) the precise definition of feature types (i.e. their attributes, associations, and behaviours) (3) common dictionaries (e.g. for physical units, coordinate references systems, physical "phenomena" definitions, etc.) and (4) maintenance procedures for definitions.

Governance also is an important control for typing granularity of features (Atkinson et al., 2004). A-priori, it is not always obvious to a designer of feature types how strongly typed they should be (see Fig. 1). In general, the more specialised the feature types, the greater will be the number required to capture the spectrum of information types used by the community.

Within the climate science community, bodies that might be expected to hold a remit for maintaining detailed Feature Type Catalogues include the UN agencies World Meteorological Organisation (WMO) and International Oceanographic Commission (IOC).

3 Climate Science Modelling Language (CSML): description

The Climate Science Modelling Language (CSML) is, to the best of our knowledge, the first attempt to apply to atmospheric and oceanographic data the full "feature-types" data integration framework described above. It is not, however, the only approach to applying the emerging ISO standards in this domain. For instance, ncML-GML (Nativi et al., 2005) extends the netCDF markup language (ncML) with ISO TC211 data models and semantics.

This section describes CSML, with its use considered in the next section.

[1] The complete work program is listed at the Technical Committees website, http://www.isotc211.org.

[2] ISO 19103 and ISO 19109 both specify that a restricted profile of UML (the object-oriented modelling language) should be used for conceptual modelling within the standards framework.

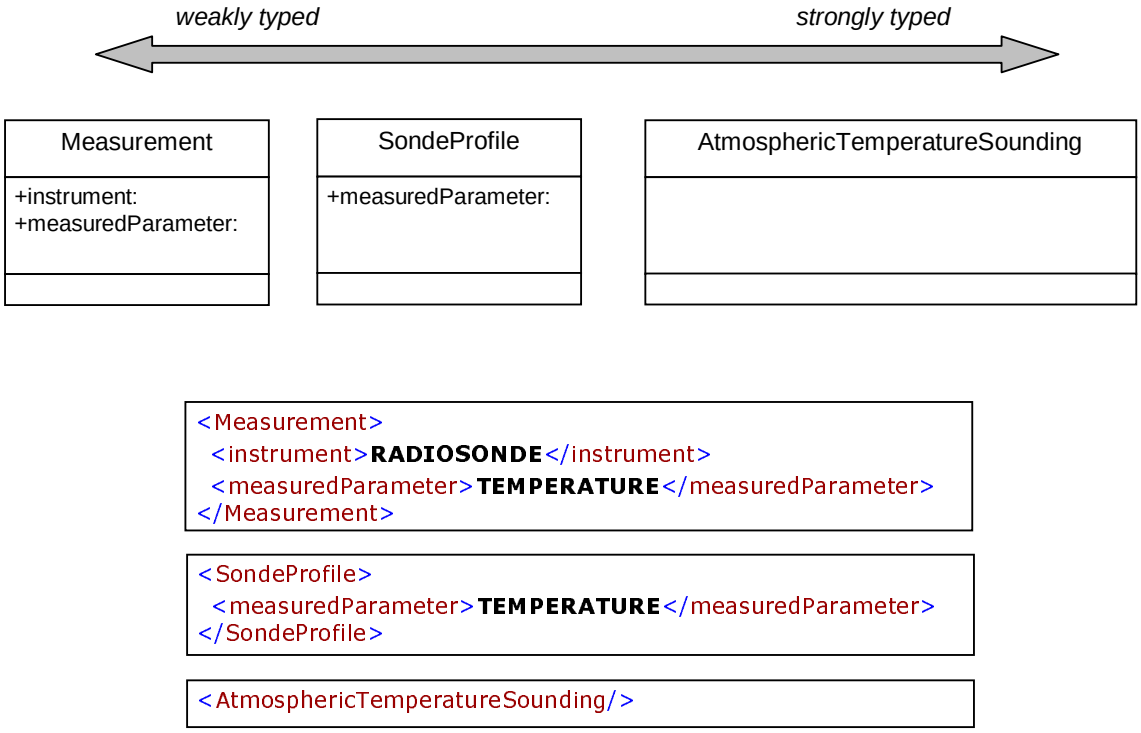

Fig. 1. Spectrum of typing granularity – examples for an atmospheric temperature sounding.

3.1 Aims

The primary aim of CSML is to provide a data integration framework for NERC DataGrid. The only scalable mechanism for this is to define a common data model across the "virtual organisation", and the only commonality across data providers is the semantics of the data. Thus, the framework offered by the emerging ISO TC211 standards – based, as it is, on formalising data semantics through "feature type" models – seems ideally suited to the NDG problem. An explicit aim of CSML is to focus on data *content* rather than the *container* in which data is held.

3.2 Design principles

As discussed earlier, governance processes supporting feature type definitions are a strong constraint on typing granularity. Given the limited remit of NDG to manage definitions on behalf of its user community (rather than the global climate science community), and given the exploratory nature of CSML, an ad-hoc target of "a handful" of feature types was an explicit aim. Nevertheless, these feature types, together, must encompass the spectrum of data types managed by core NDG partners, the British Atmospheric and Oceanographic Data Centres. A number of simplifying design principles were employed, therefore, in order to limit the number of feature types to this manageable level.

3.2.1 Offloading semantics onto parameter type

If two "features" could be regarded as the same, except for the physical 'phenomenon' of interest (temperature, salinity, wind vector, humidity, etc) then they are modelled as the same feature type. For example, both a vertical wind profile and an atmospheric temperature sounding have similar characteristics (in terms of attributes and operations performed), differing primarily in the distinguishing physical parameter (vector wind vs. temperature). Their representations are collapsed, in CSML, into the same feature type.

3.2.2 Offloading semantics onto CRS

In many cases, the geometric and topologic *structure* of data types is similar, distinguished only by the underlying "natural" coordinate system. For instance, a vertical sounding radar consists essentially of a series of measurements along a line repeated in time and directed vertically. Similarly, a scanning radar consists of a series of measurements along a directed line, repeated in time, but where now the directed line rotates around in azimuth. The underlying "natural" coordinates for the former are a fixed vertical axis in a gravity-related coordinate system together with an axis in time. The natural coordinates underlying the scanning radar are a fixed axis in a "radial" coordinate system (for backscatter measurements out from the radar) together with an azimuthal and time axis. These are regarded in CSML as the same feature

type, with considerations of the underlying coordinate system used to infer the specific data type.

3.2.3 Sensible plotting as a discriminant

The majority of data in NDG has a conventional portrayal used by practitioners for visualisation (e.g. model output is typically displayed as shaded 2-d slices, an atmospheric sounding as a line graph against height in the vertical, a marine temperature section as a 2-d contoured field against depth and ship track distance, etc.). A workable minimum granularity of feature types is determined in CSML by applying a discriminant of "sensible plotting": there should be sufficient detail within a feature type – and sufficient difference between feature types – to enable in-principle unsupervised rendering, in the conventional manner. This criterion is somewhat loose, and it remains to be seen in practice the extent to which it is satisfied with the feature types chosen. In that sense, the principle may play a more important role in evaluation than in design.

3.3 CSML feature types

On the basis of the above design principles, feature types in CSML are defined primarily on the basis of geometric and topologic structures of data. In the climate sciences, there is a strong consistency between these properties and data types identified as important by practitioners.

Such a classification enables the target to be met of a small number of feature types that will apply across the range of data curated by BADC and BODC. Seven feature types are defined in CSML, and listed in Table 1.

Of these seven feature types, the TrajectoryFeature is a pure geometry, while the other six are all discrete "coverage" subtypes (i.e. they have a spatio-temporal domain, with a value associated to each location in the domain). Thus, the essential attribute of the TrajectoryFeature is:

- **track:** a geometry providing discrete locations in space and time

while the six coverage features all have the following attributes:

- **domainSet:** the spatio-temporal geometry providing locations of values (each feature type is characterised by a different type of domainSet geometry)
- **rangeSet:** the set of values at all locations in the domainSet
- **coverageFunction:** the mapping between domainSet locations and rangeSet values
- **parameter:** the physical phenomenon for which the rangeSet provides values

Figure 2 provides some illustrations of various feature types.

3.4 CSML application schema

As well as defining seven feature types for a wide range of data in the climate sciences, CSML defines a canonical encoding, or application schema, for such data. The CSML application schema (Woolf, 2005) is an XML schema document, based on the Geography Markup Language (GML). It provides a formal "template" for climate science data to be encoded as a valid CSML Dataset. There are three basic components: (a) a collection of feature instances; (b) dictionaries for local definitions, where required, of physical units, coordinate reference systems, and physical phenomena; and (c) numerical array descriptors to provide a wrapper for file-based storage.

This structure is not strictly compliant with ISO TC211 rules for application schema (ISO 19109, 2005), but represents a pragmatic compromise until key components of a standards-based information architecture are deployed. These include referenceable dictionaries, and best-practice implementation solutions for mapping legacy data onto GML-based dataset instances.

4 CSML use

We now describe the tooling that will be developed around CSML to enable it to fulfil its role within the NDG architecture. CSML is a work in progress and much of the implementation material in this section is planned or in development.

4.1 CSML wrapper/mediator architecture

4.1.1 Array Descriptors

The wrapper mechanism for storage mentioned above amounts to being able to replace any numeric content in a CSML feature instance (e.g. for a coverage domainSet or rangeSet) with data extracted from file-based storage. This is done through an "array descriptor" which acts as a proxy for numerical content in a CSML document. A limited number of file formats are supported (e.g. NASA Ames, GRIB, netCDF). The composite pattern (Gamma et al., 1995) is employed to allow data across multiple files to be aggregated (along an existing or new dimension) in providing content for a feature instance. Thus, a timeseries of model output may be presented logically as a single GridSeriesFeature, or a collection of profile data in separate files may be logically aggregated to a single ProfileSeriesFeature.

A UML representation of the array descriptor mechanism is shown in Fig. 3. As well as providing the means for mapping and aggregating file-based data onto CSML, it supports both inline (XML) content and compact generative descriptors for regularly-spaced data. Any of these data sources (file-based, inline, implicitly generated) may be mixed in an aggregation.

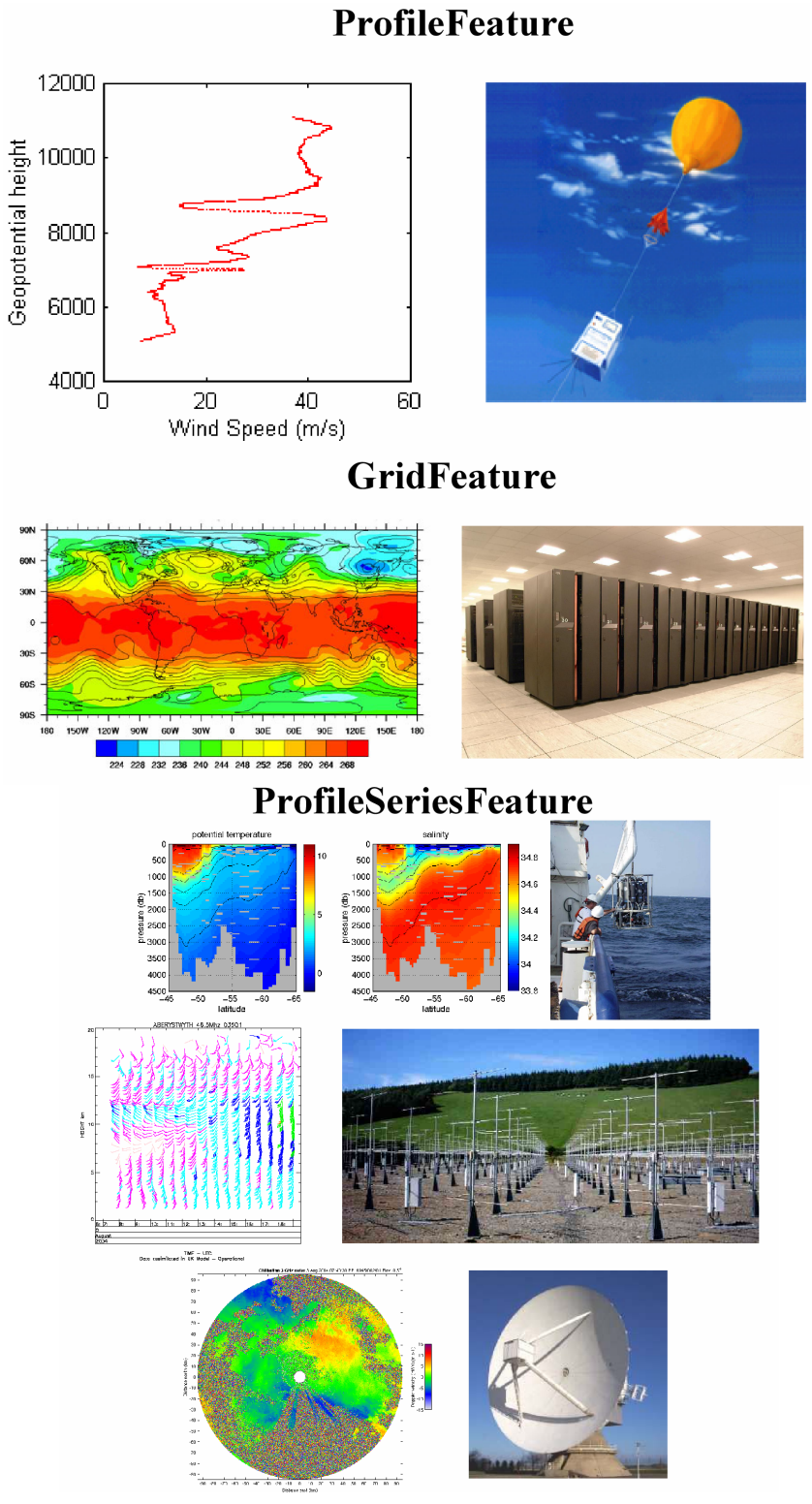

Fig. 2. Selection of CSML feature types – (a) ProfileFeature for vertical profile of wind speed, (b) GridFeature for gridded field from numerical simulation, (c) ProfileSeriesFeature for marine CTD section, vertical wind profiler timeseries, and scanning radar.

Table 1. CSML feature types.

CSML feature type	Description	Examples
TrajectoryFeature	Discrete path in time and space of a platform or instrument	ship's cruise track, aircraft's flight path
PointFeature	Single point measurement	raingauge measurement
ProfileFeature	Single profile of some parameter along a directed line in space	wind sounding, XBT, CTD, radiosonde
GridFeature	Single time-snapshot of a gridded field	gridded analysis field
PointSeriesFeature	Series of single-datum measurements	tidegauge, rainfall timeseries
ProfileSeriesFeature	Series of profile-type measurements	vertical or scanning radar, shipborne ADCP, thermistor chain timeseries
GridSeriesFeature	Timeseries of gridded parameter fields	numerical weather prediction model, ocean general circulation model

4.1.2 Parser

A general CSML parser is under development. This will demarshal a CSML document into a network of objects in-memory corresponding to the GML and CSML classes in the application schema. The parser has knowledge of the UML classes, as well as the rules used to transform the UML to XML schema (ISO 19118, 2005). Thus it can infer the mapping from an XML instance document back to individual objects. It is being implemented with SAX processing (Brownell et al., 2002) and the Builder pattern (Gamma et al., 1995) on a per-object basis. The initial parser implementation is a significant undertaking, including complete GML parsing and demarshalling code as a subset. However, as GML stabilises and increasingly respects defined UML-to-XML schema encoding rules, we expect the 'roundtripping' process to become possible to automate. In principle, a user could design an application schema using graphical UML editing tools. This could then be transformed automatically into a GML schema, with stub classes and SAX parsing code also produced automatically.

4.1.3 Services

Finally, CSML Datasets are exposed through services. While these have not yet been implemented in NDG, planned delivery services include the OPeNDAP[3] protocol and the web services of the Open Geospatial Consortium[4] (OGC; Web Feature Service, Web Coverage Service, Web Map Service), as well as file instantiation services.

In general terms, the implementation of any delivery service requires a mapping from suitable CSML feature types to the internal structures of the delivery mode in question. Instantiation of a CF-compliant netCDF file, for instance, requires a mapping to be defined from CSML feature types to CF-netCDF structures ("*variables*", "*dimensions*", "*attributes*", "*auxiliary coordinate variables*" etc.), delivery with OPeNDAP requires rules for serialisation of a CSML feature instance into OPeNDAP (*Array*, *Structure*, *Grid*, *Sequence*) data structures. Delivery via OGC web services will require a conceptually straightforward mapping – the WFS delivers feature instances directly as GML, while the WCS delivers gridded coverage data (i.e. CSML GridFeature or GridSeriesFeature) in a format of choice.

CSML supports "lazy reading" of underlying file-based data. Parsing and demarshalling of a CSML document can be done without reading data from disk. It is only at the point of delivery that the CSML "array descriptor" classes must invoke file i/o to extract the required numerical content for which they are proxy.

5 Conclusions

There is growing interest in examining the usefulness of emerging standards in geospatial data for the climate science data integration problem. Both the WMO[5] and IOC[6] have instigated activities to consider issues arising from these standards. The UK project, NERC DataGrid, is a microcosm of the more universal data integration problem – it aims to provide a unified interface to a range of environmental data across the UK and beyond. The initial focus is on data held

[3] see: http://www.opendap.org for details of OPeNDAP

[4] see: http://www.opengeospatial.org for OGC reference documents

[5] See, for example, documents and reports associated with the WMO Workshop on Metadata (http://www.wmo.ch/web/www/WDM/Workshop_metadata/documents.html), and the first meeting of the WMO Expert Team on Metadata Implementation (http://www.wmo.ch/web/www/WDM/IPET-MI-I/documents.html), Beijing, 26–29 September, 2005.

[6] See Recommendation IODE-XVIII.7 of the 18th Session of the IOC Committee on International Oceanographic Data and Information Exchange, Belgium, 26–30 April, 2005 (report available at http://ioc3.unesco.org/iode/files.php?action=viewfile\&fid=539\&fcat_id=3).

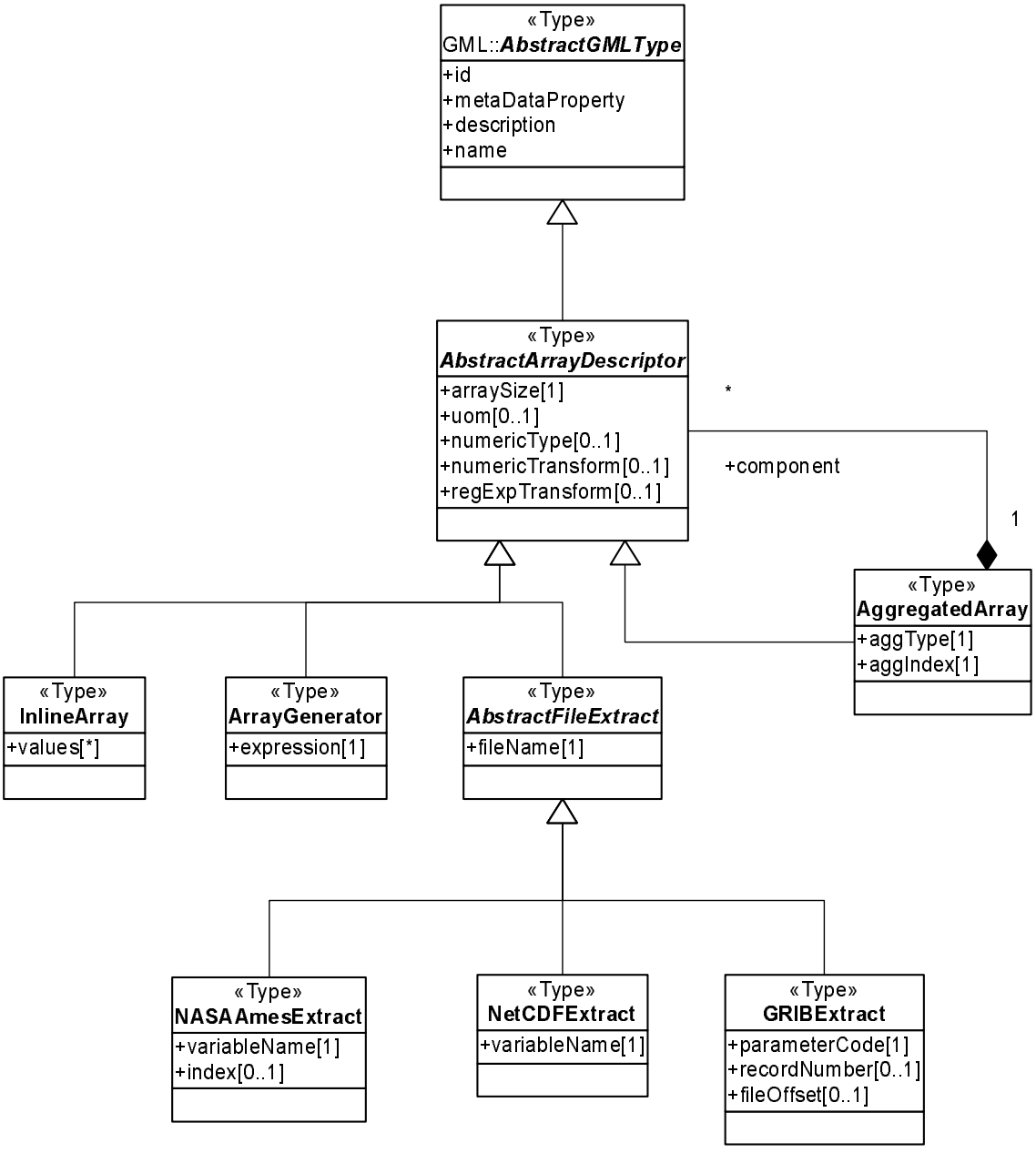

Fig. 3. UML model of wrapper for "connecting" feature instances to file artefacts.

by the British Atmospheric Data Centre and British Oceanographic Data Centre. Data integration requirements for NDG include providing a scalable architecture that leaves data in place but encapsulates the underlying heterogeneity.

The data integration framework offered by the emerging ISO standards provides just the approach needed by NDG. A series of formal data models (feature types) are constructed to characterise important data types in some application domain. An application schema then describes the structure of datasets in terms of feature types, and defines a data representation independent of underlying storage.

Governance determines the degree of granularity required in defining feature types. Consistent with NDG's remit, an initial set of seven feature types have been defined in the Climate Science Modelling Language, intended to apply across a broad spectrum of data from ocean and atmosphere. In addition, CSML defines an application schema for the logical structure of datasets.

Work is currently under way to apply CSML to some initial candidate datasets of BODC and BADC. At the same time, a parser and data delivery services are being developed for CSML. A first trial of CSML was undertaken by the EU

project MarineXML, which found it to provide a robust data integration technology (Millard et al., 2005).

The CSML effort will continue to be informed by (and inform) standards developments, and will aim to work closely with international efforts in the area.

Acknowledgements. This work was funded under the UK e-Science program through grant NER/T/S/2002/00091 from the Natural Environment Research Council.

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

References

Atkinson, R., Cox, S., Lawrence, B., and Woolf, A.: Next steps to interoperability – Mechanisms for semantic interoperability, EOGEO Workshop, University College London, 23–25 June 2004, http://y.eogeo.org/files/eogeo2004/eogeo-2004-full-atkinson.ppt, last access: 10 October 2005, 2004.

Brownell, D.: SAX2, OReilly & Associates, Inc., 2002.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J.: Design Patterns, Elements of Reusable Object-Oriented Software, Addison-Wesley, 97–106, 1995.

ISO 19101: Geographic information – Reference model, 2002.

ISO 19109: Geographic information – Rules for application schema, 2005.

ISO 19110: Geographic information – Methodology for feature cataloguing, 2005.

ISO 19118: Geographic information – Encoding, 2005.

Lawrence, B., Cramer, R., Gutierrez, M., Kleese van Dam, K., Kondapalli, S., Latham, S., Lowry, R., O'Neill, K., and Woolf, A.: The NERC DataGrid prototype, in: Proceedings of the UK e-Science All Hands Meeting 2003, ISBN 1-904425-11-9, 279–282, 2003.

Lawrence, B., Cramer, R., Gutierrez, M., Kleese van Dam, K., Kondapalli, S., Latham, S., Lowry, R., O'Neill, K., and Woolf, A.: The NERC DataGrid: Googling secure data, in: Proceedings of the UK e-Science All Hands Meeting 2004, ISBN 1-904425-21-6, 91–98, 2004.

Millard, K., Atkinson, R., Woolf, A., et al.: Using XML Technology for Marine Data Exchange, A Position Paper of the MarineXML Initiative, http://ioc3.unesco.org/marinexml/contents.php?id=28, last access: 11 October, 2005, 2005.

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 & Geosciences, 31(9), 1104–1118, 2005.

O'Neill, K., Cramer, R., Gutierrez, M., Kleese van Dam, K., Kondapalli, S., Latham, S., Lawrence, B., Lowry, R., and Woolf, A.: The metadata model of the NERC DataGrid, in: Proceedings of the UK e-Science All Hands Meeting 2003, ISBN 1-904425-11-9, 603–610, 2003.

Woolf, A., Cramer, R., Gutierrez, M., Kleese van Dam, K., Kondapalli, S., Latham, S., Lawrence, B., Lowry, R., and O'Neill, K.: Enterprise specification of the NERC DataGrid, in: Proceedings of the UK e-Science All Hands Meeting 2004, ISBN 1-904425-21-6, 294–299, 2004.

Woolf, A., Lawrence, B., Lowry, R., Kleese van Dam, K., Cramer, R., Gutierrez, M., Kondapalli, S., Latham, S., O'Neill, K., and Stephens A.: Integrating distributed climate data resources: The NERC DataGrid, in: Use of High Performance Computing in Meteorology: Proceedings of the Eleventh ECMWF Workshop, edited by: Zwiefelhofer, W. and Mozdzynski, G., World Scientific, ISBN 981-256-354-7, 215–233, 2005a.

Woolf, A., Lawrence, B., Lowry, R., Kleese van Dam, K., Cramer, R., Gutierrez, M., Kondapalli, S., Latham, S., and O'Neill, K.: Standards-based data interoperability for the climate sciences, Met. Apps., 12(1), 9–22, doi:10.1017/S1350482705001556, 2005b.

Woolf, A.: Climate Science Modelling Language Version 0.1 User's Manual, http://ndg.nerc.ac.uk/csml/UsersManual.pdf, last access: 11 October, 2005.