DATA VIRTUALISATION IN THE NERC DATAGRID

Andrew Woolf[1], Ray Cramer[3], Marta Gutierrez[2], Kerstin Kleese van Dam[1], Siva Kondapalli[3], Susan Latham[2], Bryan Lawrence[2], Roy Lowry[3], Kevin O'Neill[1].

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

Abstract:

Research in the earth sciences requires access to large and complex datasets. These data include in-situ and remote-sensed observations, and a variety of model output. Storage methods are as diverse as the data types and include numerous file formats and relational database systems. In practice, considerable effort is expended in data handling. Processing software choice is constrained by file format compatibility, and data object representations are coerced onto physical storage artefacts.

A key goal of Grid technologies is to facilitate virtualisation of resources. Essential semantic behaviour and content is abstracted from low-level implementation. Virtualisation may be applied to data, storage and computational resources. This paper presents details of the NERC DataGrid data model, and the virtualisation of earth science data it enables. The model is based upon nested hierarchies of multidimensional arrays. Standard profiles of the model are defined for important data types like 4-D gridded meteorological forecast data or oceanographic cruise measurements, for instance. Rich geo-referencing information is incorporated. An XML schema provides the mechanism for mapping physical storage artifacts onto the data objects provided by the model.

Key words: earth science, virtualisation, data model, grid, TC211, NDG.

1. INTRODUCTION

A key goal of service-oriented Grid architectures is to facilitate virtualisation of resources [1]. Heterogeneous platforms and implementations should be encapsulated behind interfaces with common syntax and semantics. Low-level resources may then be composed into higher-level services. In the case of data, details of storage location and format can then be hidden behind a uniform access mechanism [1,2].

The NERC DataGrid (NDG) project [3] is developing a Grid to provide uniform access to a wide range of environmental data held across multiple sites, and in a variety of formats. A key goal of the project is to make the transition from data discovery to use as seamless as possible, hiding from the user details of storage location and implementation. A user, having discovered a dataset of interest, then will be able to retrieve it in whole or in part from within the same grid context regardless of how and where the data is stored. Whether the data is extracted from a relational database, or aggregated from a series of flat files will be opaque to the user. Because there is a logical separation between search and discovery functionality, and data usage, the NDG separates data modelling from metadata modelling [4].

Architecting this functionality into the NDG requires a generic data model, capturing inherent structure and semantics of environmental data types, while abstracting away storage details. The result provides the means of data virtualisation and abstraction. Provided sufficient flexibility exists in the data model, it can apply across a broad range of data and a variety of disciplines. Data services and software tools which implement the model will then operate with the full range of data.

As discussed in [3] and [4] the basic NDG architecture is being developed to ensure ISO (International Organization for Standardization) compliance as far as possible as their TC211 standards are released.

2. THE NDG DATA MODEL

The NDG data model is displayed in Figure 1 using the Unified Modelling Language (UML). At the root level, a named dataset may contain both a number of parameters, and other datasets.

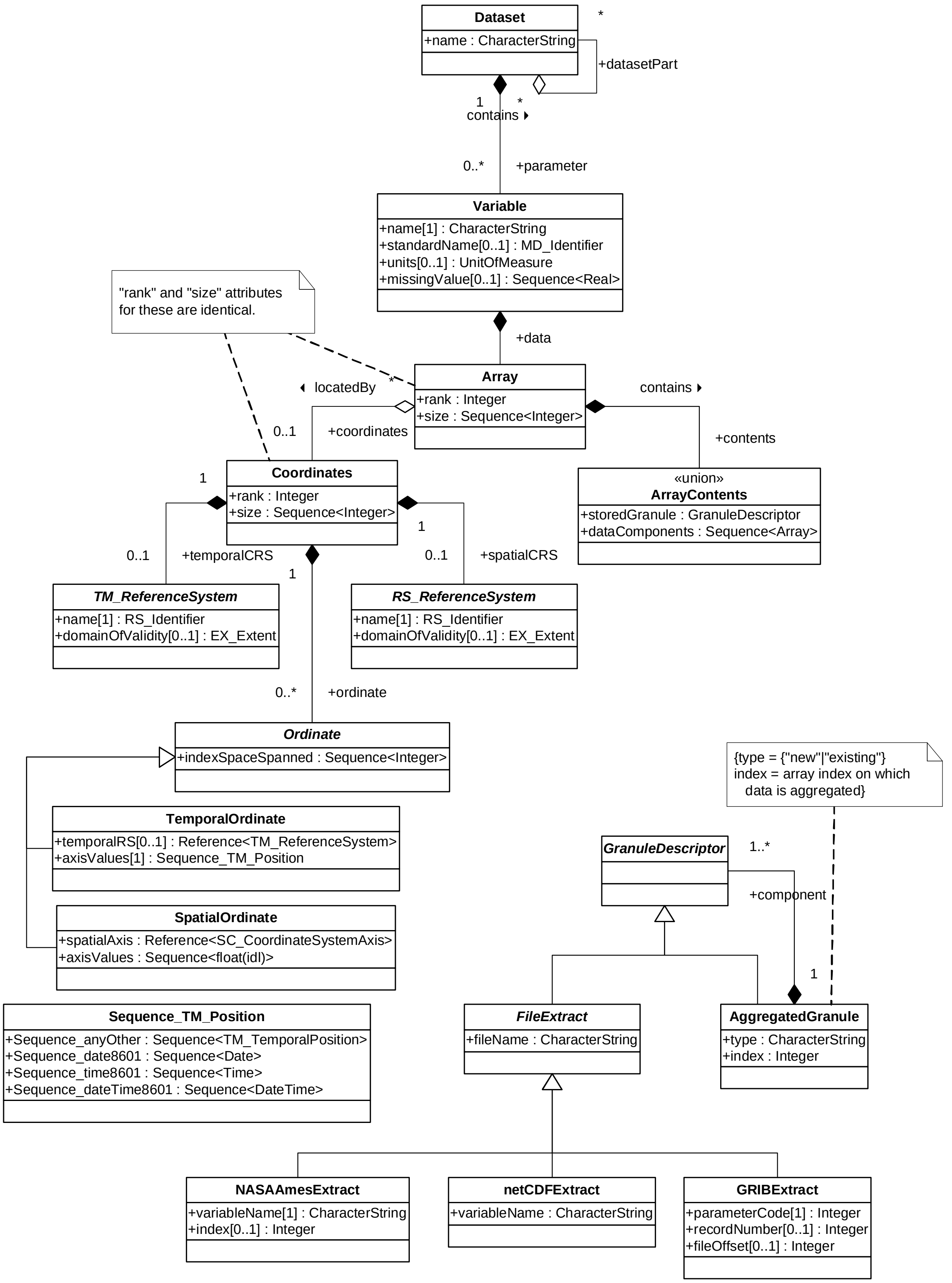
Dataset
+name : CharacterString
*
+datasetPart
1
*
contains
0..*
+parameter
Variable
+name[1] : CharacterString
+standardName[0..1] : MD_Identifier
+units[0..1] : UnitOfMeasure
+missingValue[0..1] : Sequence<Real>
"rank" and "size" attributes for these are identical.
+data
Array
+rank : Integer
+size : Sequence<Integer>
locatedBy
*
contains
+contents
0..1
+coordinates
«union»
ArrayContents
+storedGranule : GranuleDescriptor
+dataComponents : Sequence<Array>
1
Coordinates
+rank : Integer
+size : Sequence<Integer>
1
0..1
+temporalCRS
0..1
+spatialCRS
1
TM_ReferenceSystem
+name[1] : RS_Identifier
+domainOfValidity[0..1] : EX_Extent
RS_ReferenceSystem
+name[1] : RS_Identifier
+domainOfValidity[0..1] : EX_Extent
0..*
+ordinate
Ordinate
+indexSpaceSpanned : Sequence<Integer>
{type = {"new"|"existing"}
index = array index on which
data is aggregated}
TemporalOrdinate
+temporalRS[0..1] : Reference<TM_ReferenceSystem>
+axisValues[1] : Sequence_TM_Position
GranuleDescriptor
1..*
+component
SpatialOrdinate
+spatialAxis : Reference<SC_CoordinateSystemAxis>
+axisValues : Sequence<float(idl)>
1
Sequence_TM_Position
+Sequence_anyOther : Sequence<TM_TemporalPosition>
+Sequence_date8601 : Sequence<Date>
+Sequence_time8601 : Sequence<Time>
+Sequence_dateTime8601 : Sequence<DateTime>
FileExtract
+fileName : CharacterString
AggregatedGranule
+type : CharacterString
+index : Integer
NASAAmesExtract
+variableName[1] : CharacterString
+index[0..1] : Integer
netCDFExtract
+variableName : CharacterString
GRIBExtract
+parameterCode[1] : Integer
+recordNumber[0..1] : Integer
+fileOffset[0..1] : Integer

Figure 1 The NDG data model

A parameter is characterised by its name, physical units, and a numerical value indicating missing (or bad) data[1]. A standard name from a controlled vocabulary[2] may provide additional parameter type semantics (this includes the namespace authority, for example "BODC data dictionary" or "CF convention").

The parameter's data is structured as a multidimensional array characterised by its rank and size along each dimension.

The contents of an array may either be numerical data retrieved from storage or a further sequence of arrays, one per node of the parent array. This nested hierarchy of multidimensional arrays allows rich and complex data structures to be represented.

A leaf array always contains storage-derived numerical data, either from a single file, or aggregated from a sequence of component files (or further aggregations). Aggregation may be applied either along an existing array dimension, or to create a new dimension. Thus a logical four-dimensional array may be constructed from a time-series of files containing three-dimensional arrays; or two files containing respectively northern hemisphere and southern hemisphere data may be aggregated into a logical array with global coverage; or such aggregations may be combined to provide arbitrarily deep composition of file-based storage into logical multidimensional arrays.

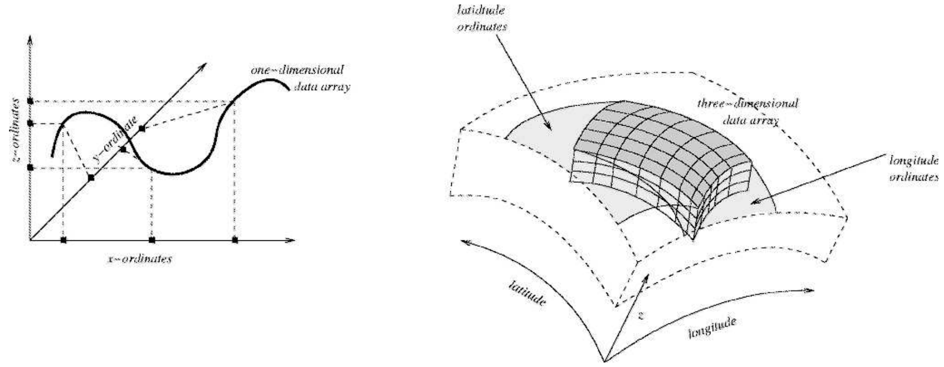

Figure 2: Coordinates in the data model. Left: one-dimensional array. Right: three-dimensional array.

Spatiotemporal location of the nodes of an array is accomplished by means of associated coordinates. These are defined with respect to spatial and temporal reference systems. The reference systems, in turn, are described in accordance with conceptual schema from the ISO 19111 and ISO 19108 standards. Individual ordinates provide values for each axis of the associated reference systems. An ordinate may span one or more dimensions of the corresponding array. Thus a one-dimensional array representing measurements along a sonde trajectory will have four associated ordinates providing measurement locations in space and time (Figure 2). Each ordinate spans the same single dimension of the one-dimensional array. A three-dimensional array from an ocean model on a rotated latitude-longitude grid will have three associated ordinates (Figure 2). Each of the latitude and longitude ordinates spans the two "horizontal" dimensions of the model array, while the depth ordinate spans the single third dimension of the model array.

The NDG Data Model significantly extends that of Woolf et al [6] in the following three respects: providing nested hierarchies of multidimensional arrays, allowing arbitrarily deep aggregation of files, and supporting much richer standards-based spatiotemporal location of data.

[1] ISO 19113 [5] provides much more sophisticated mechanisms for characterising data quality, and will be incorporated into the Data Model in the future.

[2] The MD_Identifier object is from the ISO 19115 namespace.

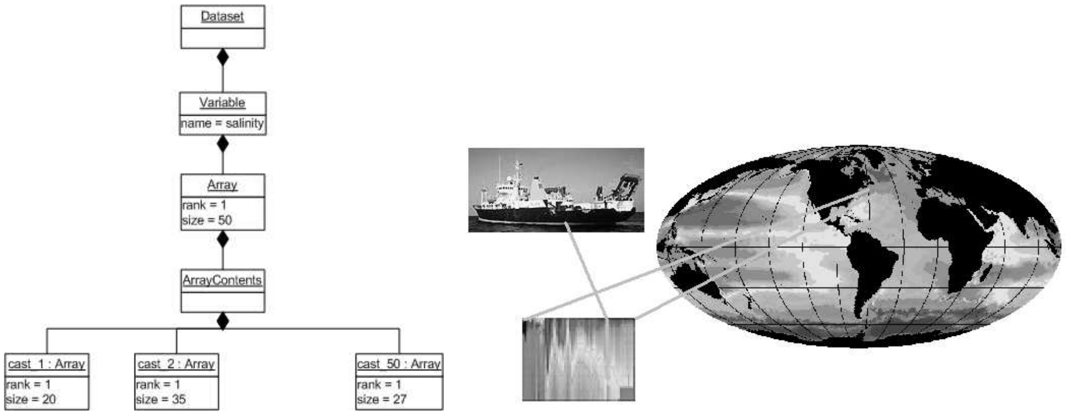

Figure 3: Data model applied to an ocean cruise.

3. APPLICATION

The generality of structures supported through nested hierarchies of arrays is broad indeed; too broad, in fact, for implementation of generic processing software. Thus the model is restricted to a small set of privileged "profiles". In practice, since this is a work in progress, the set of profiles is defined by what is implemented. Tools for processing data objects include export into a chosen file instance, and visualisation. We will describe in a more detailed report elsewhere conventions for serialising an arbitrary data object into both netCDF and OPeNDAP [7] instances. Prototype visualisation software will be limited initially to single arrays of up to four dimensions in space and time, or a two-level nesting of one-dimensional arrays. The latter may be used, for instance, to represent the set of hydrographic profiles constituting an oceanographic cruise section, as shown in Figure 3.

4. SUMMARY

An initial data model for the NERC DataGrid has been described. Based on nested hierarchies of multidimensional arrays, it describes how logical data objects are constructed from aggregations of files. Rich spatiotemporal referencing information is based on emerging international standards. The model is expected to apply across a wide range of environmental data relevant to NDG. By separating the logical structure of a dataset from the storage details, the data model provides a means of abstracting data from file locations and formats. Such virtualisation of data resources is a key goal of Grid technology.

While the model is too flexible to allow complete generic software solutions, a set of supported "profiles" of the model will grow as the project develops. As international standards from ISO are published, the data model will be adapted in accordance with their relevance. It may be, for instance, that profiles of the data model are developed corresponding to geographic feature types in the sense of ISO 19109 and ISO 19110.

REFERENCES

[1] Foster, I.,et.al, 2002: Grid Services for Distributed System Integration, Computer, 35.

[2] Chervenak, A., et.al., 2001: The Data Grid: Towards an Architecture for Distributed Management and Analysis of Large Scientific Datasets. J. Net. Comp. Apps., 23.

[3] Lawrence, B.N., et.al., 2003: The NERC DataGrid Prototype, U.K. All Hands Meeting.

[4] O'Neill, K.D., et.al., 2003: The Metadata Model of the NERC DataGrid, U.K. All Hands Meeting.

[5] ISO 19113:2002, Geographic Information-Quality Principles.

[6] Woolf, A., et.al.,2003: A Web Service Model for Climate Data Access on the Grid. Int. J. HPC. Apps. 17 (3)

[7]Distributed Oceanographic Data System, http://www.unidata.ucar.edu/packages/dods/index.html