MINUTES OF GGF8 DFDL WG MEETING

Cedar Room, 8:00-10:30am, 25th June 2003
Chairs: Martin Westhead, Alan Chappell
Present: approx 23

1. Introduction and Agenda
2. Grounding in Existing Work
3. DFDL Background
4. Example Repository

1. Introduction and Agenda (Martin Westhead)

Martin presented the GGF IPR policy document and sign-up sheet was initiated. He then presented a brief introduction to the motivation and aims of the work of the group.

2. Grounding in Existing Work (Alan Chappell and Martin Westhead)

Alan gave a presentation entitled Describing "Arbitrary" Data using XML which described his work with Binary Format Description (BFD) as an internal research project at PNNL. The presentation showed how BFD, as an extension of Extensible Scientific Interchange Language (XSIL), uses an XML description to help interpret the data in the file and how one description can represent a whole class of data.

- Comment: Be aware of similar initiative, Open Data Access Protocol

- Question: where do you put structure information like that in HDF4 and HDF5?
- Comment: HDF5: This describes the format

Alan gave examples of places where BFD is used e.g. the Scientific Annotation Middleware project.

- Question: Do you provide a schema?
- Answer: Description is a type of schema for the data. BFD provides a schema for the language.

- Comment (Jim Myers): XML type description more "trusted" than having arbitrary Java code running on server in web service environment.

Martin (standing in for Stephen Rutherford) gave a brief summary of BinX and highlighted the differences between it and BFD. BINX being developed as part of the eDIKT project and also uses an XML description of what the data contains. BinX is not based around XSIL and is mainly geared at binary data. Currently, BinX being used with the AstroGrid project

3. DFDL Background (Martin Westhead)

Martin gave an overview of the proposed DFDL and its uses and benefits. WG effort will entail a general structural description, an XML representation, and an ontology definition. Martin discussed a "dense" document on the web site specifying the structural description. This document is a draft. He asked for comments on the draft.

- Question: Is the binary data written and read using the language?
- Answer: No, we're trying to be format agnostic. Can describe and read legacy data.

- Question: Can describe Cobol?
- Answer: Should be able to.

DFDL does not express semantics, but provides semantic labels instead. The meaning attached to the arbitrary labels is in the underlying ontology.

- Comment: challenge in ontology is in the relationship between the labels.

- Question (Reagan Moore): Have you looked at the set of relationships that you need to deal with the ontologies? There aren't that many.
-Answer: Some standard relationships will be represented in the default ontologies produced as part of the WG effort. Users create their own extensions to include more specific relationships.

- Question: Re: CSV example in presentation, where are the names that are associated with the different bits of data, different types?
Answer: That is in the ontologies. The ontology would need to be able to describe the minimum "char". Can do something richer in the ontology to enable a more powerful manipulation. Need to define new types and operations to work with them, if they are not primitives (extensions).

- Comment: Can enable navigation of the data through the description using the library

- Question: What APIs would be provided by DFDL to access the data described. –Answer: API's developed through WG process, but will likely include the following

some form of reflection
some navigational / traversal ability
a way of extracting data represented

indexing ability

- Question: Lots of discussion about XPath etc. So you describe your data in DFDL, and can go back to that data without having to request the whole thing?
- Answer: Yes, treat it as if was in XML format to pull out specific elements of interest. Can be used for transforming, sub-setting, or integrating data.

- Discussion: Lots of questions about implementations. To do this, need to understand the file descriptions. Issues raised included:
"extinct" formats - DFDL would be a good tool to resurrect them;
maintaining integrity between definitions and file contents when content is volatile - could use self-generating documentation;
validation of data file against DFDL description. When, where, how?

Where are we going next? Push ahead and produce strawman of XML representation and of an ontology. Martin gave a description of the WG's proposed goals. One was to build a repository of good examples to test on.

4. Example Repository (Alan Chappell)

WG effort will benefit from having repository of real data to explore and test with. Need the community to supply data and information about that data. The data needs to be insensitive for public access. Need to work to provide descriptions to find out how to move forward the work. Request for data and assistance.

Bruce Barkstrom gave a description of some of the data examples held at the NASA sites: energy flow into and out of the atmosphere. These could have variant structures and parameters kept in different files. He also highlighted the National Digital Information Infrastructure Preservation Project undertaken by the Library of Congress. This brought up the issue of how to check if the new version of the data is identical to the old one. He also mentioned the usefulness of being able to reference particular labels within files, e.g. all pixels representing a hurricane.

- Comment (Reagan Moore): Number of other similar projects that need to be reviewed for relevance to DFDL;
Common Warehouse Metadata from OMG
Binary format data project with W3C to provide "denser" version of XML
ASN1 in the telecoms industry

- Comment (Reagan Moore): Applications and the operations they can perform; Data descriptions and their structures, must be a matching for interoperable use.

Martin noted the new mailing list dfdl-wg@gridforum.org.