Access to Data in Files Working Group

Global Grid Forum, Data Area

Administrative Information

Name and Acronym:
Access to Data in Files Working Group (ADF-WG)

Chairs:
Brian Collins, collinsb@uk.ibm.com
Neil Chue Hong, n.chuehong@epcc.ed.ac.uk

Secretary(s)/Webmaster(s):
tbc, tbc@tbc

Email list:
adf-wg@ggf.org (to be confirmed)

Web page:
http://forge.ggf.org/projects/adf-wg/ (to be confirmed)

Charter

Focus/Purpose
Data access and integration activities relating to the Grid have focused on applications where data is stored in relational databases and XML repositories. However, in many scientific and commercial domains a huge amount of data is stored in large numbers of files, where the internal structure is typically application specific. The group seeks to promote standards for the development of grid file data access services, focusing principally on providing access to the content of files which is consistent with that for relational databases and XML repositories. In particular, the group does not seek to develop new file storage systems, nor file definition languages or formats, but rather to make such systems more readily usable individually or collectively within a Grid framework and interoperable with other forms of data resources described by the DAIS-WG.

Scope
The ADF-WG will focus on the specification of Grid File Data Access Services. The specification will provide features including the query and update of virtualized files. It is anticipated that multiple interfaces could be recommended for access to data in files, for example the use of SQL and/or XQuery to query and update files.

In order to produce the specification, the ADF-WG will identify and document requirements and usage scenarios to facilitate the discussion for the standardization. Existing surveys of the state-of-art technologies and specifications will be used to inform the process.

Goals
1. Informational Document: **"Access to Data in Files: Requirements and Usage Scenarios"**
   This document will describe the requirements and usage scenarios to facilitate the discussion for the standardization.
   a) Draft by GGF 12
   b) Final by GGF 13

2. Recommendation Document: **"File Data Access Services Specification"**
   This document will produce a specification for services which provide access to data in files. These services will provide operations which are consistent with those recommended for relational database and XML repositories (in effect producing a Files Realisation for Data Access Services, in the language of the DAIS-WG). They will also be consistent with those recommended for Grid File Systems and other common services
   a) Draft by GGF 13
   b) Updated draft by GGF 14
   c) Final by GGF 15

Management Issues

Evidence of commitments to carry out WG tasks
Brian Collins has worked on the UK eScience OGSA-DAI reference implementation and the UK eScience eDiaMoND (Digital Mammography) grid prototype. He is currently an editor for the DAIS – Relational Realization. He can commit the time required for this WG supported by his employer.
Neil Chue Hong is the Project Manager for the second phase of the OGSA-DAI project. He has previously worked on the definition of file access and query prototypes for OGSA-DAI, and is working with several projects to understand the requirements for access to data in files for incorporation into the OGSA-DAI software. He is currently Secretary for the Data Transport WG. His time is currently funded by the DAIT project.

Pre-existing Document(s) (if any)
Martin Westhead, *Data Access and Integration Services – Files Realization*, DAIS-WG, GGF9, 2003
Reagan Moore, *Operations for Access, Management, and Transport at Remote Sites*, DT-RG, Jan 2004.
Vijayshankar Raman, *Services for Data Access and Data Processing on Grids*, GFD-I.14, February 2003.

Document from Ohio State University (to be confirmed)

Relevant standards include:
- WebDAV/HTTP
- FTP
- GridFTP
- POSIX
- NFS/AFS

Exit Strategy
The WG will be finished after submitting the recommendation document File Data Access Services Specification.

Potential group lifetime extensions could cover access to data in streams, or extension to the File Data Access Services Specification to cover management of data in files.

Any other relevant information

Evaluation Criteria (from GFD-C.3)

Is the scope of the proposed group sufficiently focused?

The group will focus on Grid Services which are applied to access of data stored in files. The goal is to produce a specification of File Data Access Services.

Are the topics that the group plans to address clear and relevant for the Grid research, development, industrial, implementation, and/or application user community?
Yes. Grid research projects in US and Europe have worked on the topic, and provided prototype implementations. There is a strong need to have standard service interfaces for access to files in the grid which are consistent with those for relational databases and XML repositories. The recommendations by this WG will facilitate further implementations and usage in industry and user communities.

Will the formation of the group foster (consensus–based) work that would not be done otherwise?
Yes. There are several research projects that prototype the functionalities of the topic to be recommended by this working group. The formation of the group will foster standards that are greater than the work done by any single group. Proposed co-chairs (IBM and the University of Edinburgh) and potential participants (Ohio State University, Queens University Belfast) are from organisations that have different projects at institutions across various regions of the world to implement File Data Access Services. These recommendations could potentially used in US (by projects funded by multiple funding agencies like NSF, NIH, DoE, etc), Europe (UK eScience) and Asia-Pacific. There has been also been interest from other groups working from Europe (EGEE).

Do the group's activities overlap inappropriately with those of another GGF group or to a group active in another organization such as IETF or W3C?

No. There are several groups with relevant, complementary work and we will be careful, whereever possible to present standards consistent with theirs:

 - The ADF WG will support the naming mechanism defined by the GFS WG
 - The ADF WG will supply an interface consistent with that proposed in the DAIS WG

We will endeavour to work with the file replication group, the storage group and the data format group to best leverage the work that they are doing.

In addition our aim is to make maximum use of existing standards material and to draw on, and be consisted with the interfaces and semantics of as many existing file access technologies as possible.

Are there sufficient interest and expertise in the group's topic, with at least several people willing to expend the effort that is likely to produce significant results over time?
Yes. This section to be updated once call for interest has been distributed. We also intend to involve user communities, researchers and industry partners in the generation of the requirements and usage scenarios informational documents.

Does a base of interested consumers (e.g., application developers, Grid system implementers, industry partners, end-users) appear to exist for the planned work?
Yes. The output of the WG will be applicable to every Grid application that wants to have access to files to facilitate Data Virtualization. The interested consumers could also include companies and organisations like IBM and UK eScience projects, and especially relevant to projects and technology transfer partners who have a requirement to access data in files produced by legacy applications.

Does the GGF have a reasonable role to play in the determination of the technology?
Yes. The GGF is the right place for this work because we are going to provide a standard specification for services which allow access to data in files.