Grid Forum Research Group Charter

Research Group Name:

Preservation Environments

The "Preservation Environments" research group acronym is "PERG".

Chairs:

Reagan W. Moore, moore@sdsc.edu
Bruce Barkstrom, b.r.barkstrom@larc.nasa.gov

Interim secretary:

Andre Merzky, merzky@zib.de

Mailing List:

General Discussion: data-wg@sdsc.edu
To subscribe: Majordomo@gridforum.org with the text
subscribe pe-rg "your name" <"your E-mail address">
Archive: http://www.gridforum.org/

The Preservation Environments Research Group web site will be linked from http://www.gridforum.org/. The web site will contain copies of all working drafts, lists of working group participants, and lists of current working group activities.

Description of Research Group:

The Preservation Environments Research Group of the Grid Forum will develop a GGF Informational Document for Persistent Archives based on virtual data grid technology. The Information Document will demonstrate

- Significant implementation
- Successful operational experience
- Widespread use

of persistent archive technology. The Information Document on Persistent Archive Concepts, submitted to GGF from the Persistent Archive Research Group, will be used to guide at least three different implementations of persistent archive technology. Comparisons will be made of the features provided by each environment, and how these features can be supported by Data Grid technology. We will work with the InterPARES project (International Research on Permanent Authentic Records in Electronic Systems), several national archives that are associated with InterPARES such as Australia, Canada, Singapore, and NARA, and research groups that are building testbeds such as

the UK Data Curation Center, the National Historical Publications and Records Commission (NHPRC) persistent archive testbed, and the US Library of Congress National Digital Information Infrastructure and Preservation Program.

The proposed Information Document will serve two purposes:

- Describe to the preservation community best practices for implementing distributed preservation systems
- Identify for the grid community the consistency constraints that are required between grid services for a viable preservation system.

Preservation systems are already being implemented that contain components based on data grid technology. Examples include:

- Computing Centre of Academia Sinica, Taiwan
- NASA Langley Research Center
- US National Science Foundation, National Science Digital Library persistent archive. The NSDL persistent archive maintains persistent copies of material retrieved from web crawls, housed in a data grid implemented at the San Diego Supercomputer Center using the Storage Resource Broker. The system integrates digital library metadata access mechanisms (Open Archives Initiative) and GSI Grid Security Infrastructure.
- US National Archives and Records Administration prototype persistent archive. The system is distributed between NARA, University of Maryland, and the San Diego Supercomputer Center. A major development effort is support for peer-to-peer federation of independent registry catalogs between the sites.

Multiple sites have indicated interest in the implementation of preservation systems. They include:

- US National Historical Publications and Records Commission. Proposals have been submitted to the NHPRC to implement 5 persistent state archives that use data grid technology to automate archival processes.
- US Library of Congress National Digital Information Infrastructure and Preservation Program. The NDIIPP is seeking multiple implementations of preservation systems to test interoperability and the ability to migrate preserved collections between systems.
- Japan National Institute of Advanced Industrial Science and Technology, Grid Technology Research Center. A preservation system is planned under the Tsukuba WAN environment.

Deliverables:

Document describing implementation approaches for preservation environments from at least three different groups and best practices for the implementation of a persistent archive based upon virtual data grid technology

Goals/Milestones:

- GGF10 (March 7-10 2004) Create Recommendations document, specifying which capabilities must be provided
  - Identify at least 3 implementations
- GGF11 (June 6-9 2004)
  - Start report on which features have been implemented
- GGF12 (October 2004)
  - First draft of implementations
  - Discuss missing features
- GGF13
  - Production persistent archives
  - Recommendation draft
- GGF14
  - Final draft

Potential Document Committee Participants:

- Bruce Barkstrom - LARC, b.r.barkstrom@larc.nasa.gov
  Interest is preservation of massive Earth Sciences collections, including the tracking of provenance information for derived data products
- Reagan W. Moore – SDSC, moore@sdsc.edu
  Interest is persistent archives used that manage technology evolution, and identification of the overlapping infrastructure components with Data Grids and digital libraries
- Sam Sun – CNRI, ssun@cnri.reston.va.us
  Interest is the use of persistent handles to identify digital entities, and how these would be implemented within persistent archives
- Isao Kojima – AIST, kojima@ni.aist.go.jp
  As part of the Japan National Institute of Advanced Industrial Science and Technology, Grid Technology Research Center, a preservation system is planned under the Tsukuba WAN environment.
- Jane Xu – IBM, jxu@us.ibm.com
  Interest is integration of preservation systems into the IBM data grid
- Kerstin Kleese K.Kleese@dl.ac.uk
  Interest is use of data grid technology for preservation
- Andre Merzky – ZIB, merzky@zib.de

Interest is use of data grid technology for preservation

Last Updated 29 January 2004