Grid-based Search and Data Mining Using Cheshire3

Ray R. Larson, UC Berkeley, USA

azaroth@liverpool.ac.uk

Robert Sanderson, University of Liverpool, UK

ray@sherlock.sims.berkeley.edu

This presentation will describe our recent research in designing and developing grid-based search and data mining facilities for digital library services. An important aspect of this research is concerned with providing effective and scalable IR services for digital libraries as these diverse collections grow to sizes measured in terabytes and petabytes. The Cheshire project has had a central research focus on large-scale digital library collections for more than a decade, with a current focus on supporting distributed digital libraries in a Grid evironment. At the same time we have been
prototyping systems for very long-term digital preservation, and examining how grid-scale information retrieval systems can interoperate with petabytes of diverse data stored over many years.

Critical functionality for Digital Libraries involves massive, secure, storage and effective information retrieval capabilities to search over the content stored. To provide such capabilities effectively, there must be a flexible and extensible series of “Grid Services” with identifiable objects and a known API's to handle the IR functions needed for Digital Libraries or other retrieval tasks. The Cheshire3
system builds on the work of the Cheshire project over the past decade to define and implement an easy to use set of IR objects with precisely defined roles that can effectively provide a Grid Service for IR. We will discuss how distributed storage technologies like the SRB and IRODS are being used in Cheshire3, and the issues of efficiency in such a computing environment for large-scale demonstration projects such as the National Archives and Records Administration Preservation prototype.

This work is a collaboration between the University of California, Berkeley, the University of Liverpool and the San Diego Supercomputer Center.