The Adaptive Viral Archive

B. Barkstrom[1]

[1]NASA Langley Atmospheric Sciences Data Center, US
Bruce.R.Barkstrom@nasa.gov

Abstract. In this paper, the author considers how semantic technologies may impact the design of future archives, which require both very secure approaches to maintaining data, metadata, and their provenance, as well as the most cost effective operation possible. One potential application for semantic technologies may be improving the automation of system configuration, allowing an archive to be "self-replicating" or "viral" – meaning that the archive infrastructure as well as its data could automatically pack itself into a self-contained structure that could be transmitted to a new site and automatically unpack itself. A more challenging problem is likely to be dealing with the evolution of the semantics of the collection and of the archive's user communities over an extended period of time.

Keywords. Self-replicating archives; Ontology Evolution

Semantic Technologies and the Archive of the Future

A decade's experience of working with a Petabyte archive of binary scientific data suggests the need for a substantial improvement in the archival community's ability to deal with the evolution of archival system technology and of the data access patterns of the user communities that access these data. Of particular concern is the need for replication of very large amounts of data and metadata while maintaining careful controls on the provenance of these items.

In this paper, the author will explore an architecture based on the Open Archives Information System (OAIS) Reference Model, in which the design of the archive has the potential for highly automated replication of both the archive's data and the infrastructure of the archive itself. Such an approach would allow an archive to assume a "viral" nature, in which it could automatically package itself together with the instructions for unpacking and installing itself. Replication of data can be handled in the same way. This means that an archive could be "passivated", sent to a new location, and then "reactivated" with minimal human intervention. Some human control should probably still be required to ensure that the replication was authorized to use the resources on which the new archive was being instnatiated and to ensure the protection of intellectual property rights of the data being replicated.

Semantic web and semantic grid technologies may be useful for dealing with the complexities of automated configuration management for installing a new archive. Thus, in the foreseeable future we may expect the instructions for packing and unpacking the archive to be embedded within XML protocols that have semantic elements. These elements would inform the unpacking procedures about the key components that are sensitive to the infrastructure in which the archive software will reside.

However, a more difficult problem is likely to be supplying semantic components to deal with the evolution of the environment, including the diversity of the user community linguistic dialects. In the case of the Atmospheric Sciences Data Center, there are at least five distinct communities of users. Each community has some vocabulary elements in common and some that are unique to that community. As a result, the linguistic model that supports user searches needs to become adaptive in order to deal with this diversity.

Dagstuhl Seminar Proceedings 05271
Semantic Grid: The Convergence of Technologies
http://drops.dagstuhl.de/opus/volltexte/2005/389

As part of our work in trying to improve user access to our data, we have been exploring a somewhat abstract description of the ways users search for the data (and services) that interest them. The basic organization of the data within our archive can be viewed as a hierarchical classification of files, which suggests a tree- structured search approach as a possible foundation. In the classification-based search, each selection by a user in traversing the tree is equivalent to selecting a parse string from a strongly constrained language. At the same time, the tokens themselves are not necessarily strings of characters, meaning that an adaptive user interface might present alternative symbols suited to a particular user subcommunity. To put it slightly differently, the representation of the tokens may alias character strings (or images) that the search interface disambiguates as it interacts with the user.

This formal idea becomes more interesting when we allow users to engage in multi-faceted searches, in which their search strategies are no longer confined to traversals through the file classification tree. Based on our current experience, we expect different user communities to start from different entry points in an adaptive interface and to follow different paths through the search nodes of the site. We believe it is likely to be particularly useful to categorize users by their search paths and to adapt the presentation of potential paths to these communities.

While this classification appears to open substantial improvements in the user search experience, it also creates at least two interesting complications to semantic operations. First, it is clear that web sites will need to accommodate different ontologies for different user communities. Because different communities overload key phrases with different meanings and contexts, some form of automatic translation between dialects appears to be necessary. At the same time, automatic translation still seems to be a fairly difficult business. Semi-automated translation is expensive and slow. Second, it is also clear that user dialects (and data world views) evolve over time. This means that key phrases and ontologies will need to find ways in which they can be maintained as the community's usage of them changes over time.