Society of Archivists Data Standards Group

A Guide to Archival and Related Standards
Standards applicable to archives; for the digital delivery of repository guides, finding-aids, and images of material from collections.

Title
OAI-PMH - The Open Archives Initiative Protocol for Metadata Harvesting

Name of Standards Developing Organisation
OAI - The Open Archives Initiative

Current version
Version 2.0, June 2002

Replaces
Version 1.1, July 2001

Abstract
OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting) is a standard used to harvest (collect) metadata descriptions of digital information so that services can be built using metadata from different sources.

Description
OAI (The Open Archives Initiative) develops and promotes interoperability standards that aim to facilitate the dissemination of content. OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting) provides an application-independent interoperability framework based on metadata harvesting.

Within the OAI-PMH framework are Data Providers, who administer systems that expose their metadata through the OAI-PMH, and Service Providers, who use the metadata harvested via the OAI-PMH as a basis for building services.

Metadata refers to data that is about resources, such as archive materials, bibliographic resources or research papers. Harvesting refers to the process of collecting metadata. A harvester issues a request to a provider as a means to collecting metadata from their repository. The service provider generally aggregates metadata from different data sets to provide users with the benefit of a cross-searching service.

It is important to understand that the word 'archive' in the context of OAI-PMH is not used to denote collections in the sense that archivists understand, but
rather it refers to digital information, and in particular research information, such as e-prints.

A service, such as a library catalogue, may wish to provide information to users that is stored in other repositories. The metadata harvester is pointed at the repositories to gather the information and integrate it into the library catalogue. The metadata is typically provided in Dublin Core, which is a standard set of 15 metadata elements (title, creator, subject, description, etc.). The catalogue then includes the metadata and the user can be linked from this descriptive information to the content, which is still at the original repository’s site. Because OAI-PMH harvests metadata, not the actual content, any possible limitations on access to the content will not be violated.

One of the largest examples of a service using OAI-PMH to harvest from different providers is OAIster, based at the University of Michigan. OAIster harvests metadata from over 1,000 contributors. All the contributors, as data providers, must provide an OAI-PMH interface so that their metadata can be harvested. Google Scholar is also an OAI-PMH service, indexing many types of free and commercial publications. The Archives Hub has an OAI-PMH interface, which means the Hub descriptions are exposed for harvesting, enabling other services to gather the metadata. By default OAI-PMH enables all records to be harvested and enables selection by date (so that only recently added records can be harvested). To harvest other groups of records, specific record sets need to be created.

OAI is linked conceptually to the idea of open access in scholarly publishing and the creation of institutional repositories. For this reason, it has garnered considerable support from information and library science professionals. It is based on common standards - HTTP and XML - and is intended as a low-barrier solution to interoperability. However, it is not always as straightforward as might appear to be – often a level of manual intervention in harvesting metadata is required and incremental harvesting may be problematic. There is a high level of flexibility in the protocol and in Dublin Core, which can create obstacles for service providers. Data providers may only provide a minimal level of information (for example, only author and title) because there are no elements that are compulsory to include in the metadata. It may not, therefore, be possible to build sophisticated search services from the aggregated metadata.

Once a data provider has made their service OAI-PMH compatible they need to validate, to make sure that it follows the standard, and then report the 'archive' to the OAI. They will then be placed on the OAI list of data providers where service providers can find services.

The Open Archives Initiative Protocol for Metadata Harvesting has become an established and widely adopted international standard. By utilizing standards to organise information, the Protocol has successfully helped to increase access to digital resources.
Next month
Next month we will look at the information technology security management standard ISO/IEC 27001:2005, which establishes the requirements for initiation, establishment, implementation, maintenance and improvement of an ISMS within and organisation

Jane Stevenson
Archives Hub