Open Repositories 2.0:
Harvesting Community Annotations to Enhance Discovery Services

Jane Hunter, Imran Khan, Ron Chernich and Anna Gerber
The University of Queensland
Australia

Introduction

Over the past few years, collaborative social tagging and annotation systems that involve communities of users creating and sharing their own metadata, have exploded on the Internet. Examples of such systems include: Flickr, Del.icio.us, Connotea, YouTube, LastFm. Such systems are exemplary of the Web 2.0 phenomena because they use the Internet to harness collective intelligence. Although there are issues associated with the quality of the metadata generated by online communities, there are also significant advantages including the cost benefits of leveraging community effort to generate metadata and enhanced search and discovery services that result from richer, more relevant metadata and rankings of resources.

In this paper we describe the HarvANA (Harvesting and Aggregating Networked Annotations) system that we are developing at the University of Queensland. The objective of HarvANA is to develop an efficient streamlined system (that is based on open standards and comprises a set of open source services) that can leverage the explosion of community annotation/tagging systems and exploit the resulting metadata to improve discovery and reasoning across open repositories.

Within the HarvANA system, community annotations are stored on (one or more) Annotea-compliant annotation servers that are separate from the collections that they are annotating. An OAI-PMH interface has been built on top of the Annotation server(s). This enables the periodic harvesting of new annotations (since the last harvest) by sending OAI-PMH (HTTP) requests to the server(s). The harvested annotations are then aggregated with the institutional metadata (IM), to enrich the metadata store with community knowledge. Figure 1 provides a high-level view of the HarvANA system architecture.

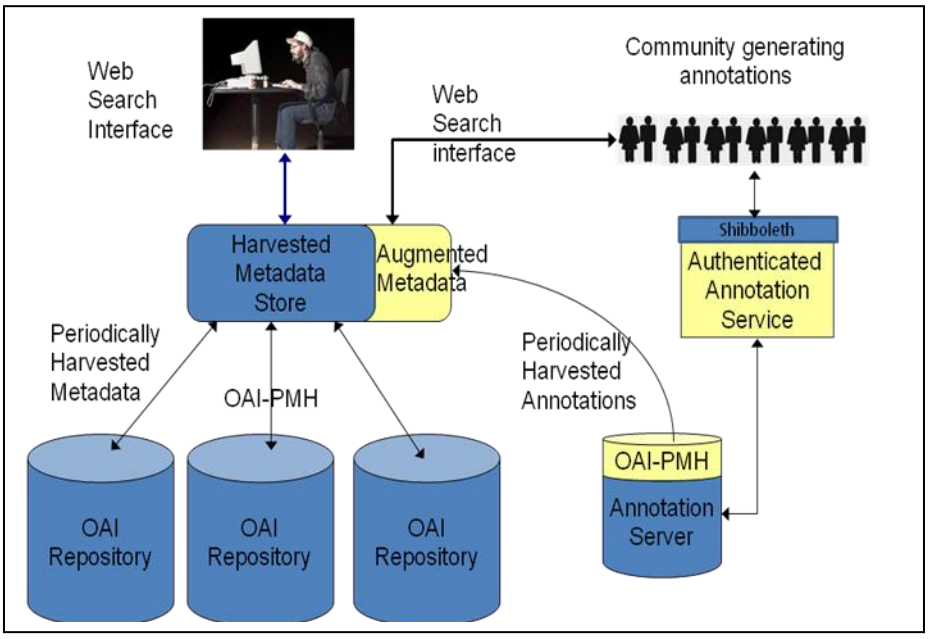
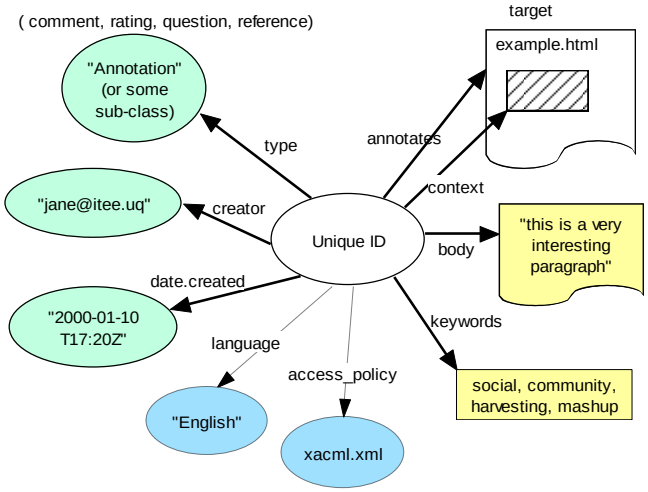


Figure 1: HarvANA system Architecture

Underlying Technologies and Functionality

HarvANA uses the W3C's Annotea annotation protocol and an RDF Jena data store for storing and querying annotations. In addition, a security interface based on Shibboleth user authentication and XACL access controls has been implemented which restricts access to the annotation server (and individual annotations) to members of specific online communities.



An Annotea client plug-in for Internet Explorer and Firefox has been developed that enables users to create and attach annotations to resources retrieved via a Web Search Interface. The system supports the annotation of web pages, images, video, audio and 3D objects (protein crystallography structures). In addition, the system provides a user interface for browsing and searching annotations. Figure 2 illustrates the extensible Annotea-compliant RDF model/schema that is used to record annotations/tags and their associated metadata. Users can search across metadata fields including: creator, date, keywords or free-text searching over the description. Quality control of the community annotations is assured by validating annotations/tags against the schema and restricting tags to a controlled vocabulary or ontology – accessible via pull-down menus within the annotation creation interface.

The Annotea server is implemented using a Tomcat Java Servlet. The RDF annotations are stored using the Jena API over a MySQL database. The OAI-PMH interface on the Annotea server was developed by mapping the RDF annotations to unqualified Dublin Core and incorporating OCLC's OAICat Java servlet within the Tomcat Java Servlet container. This enables HTTP requests to be periodically sent to the Annotea Server to retrieve any new or updated annotations as XML records. These can be incorporated within the original institutional metadata store to enhance the search and discovery service.

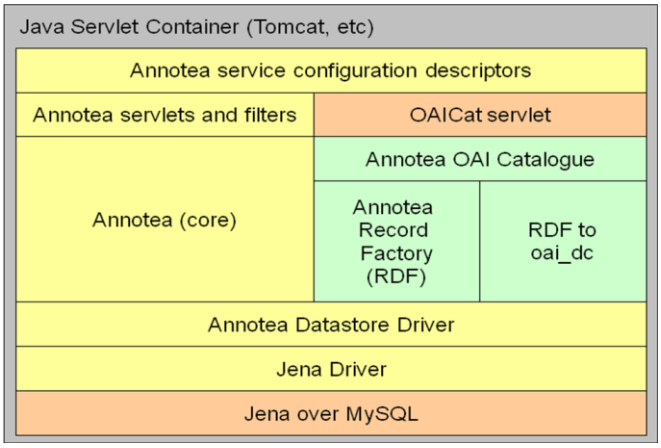


Figure 3: Annotea OAI-PMH Implementation

An Architectural Case Study and Test Bed

We are currently evaluating the HarvANA system through a collaboration with the National Library of Australia (NLA). Our aim was to assess HarvANA as a value-add community service that can run in conjunction with existing repository search services, such as PictureAustralia, MusicAustralia or PeopleAustralia. PictureAustralia is an NLA project that provides a federated discovery service to more than 1 million images from 31 contributing organizations. PictureAustralia's Web-based search interface uses a central database of metadata held at the NLA, that has been harvested from the contributing organizations using the OAI Protocol for Metadata Harvesting (OAI-PMH). Incremental OAI harvests (of updates) are carried out on the larger sites every night and the smaller sites, once per week.

To evaluate HarvANA we acquired a selection of architectural images and metadata from PictureAustralia. We built a local replica of the PictureAustralia system on a MySQL database and Web server at the University of Queensland. We then developed an (OWL) ontology of architectural terms that limits the keywords or tags to a set of controlled, machine-processable terms. The generic annotation creation interface was customized by tailoring the underlying annotation schema and incorporating the architectural ontology. We then set up an annotation server for storing the annotations and set about creating annotations about the images and storing them on the server. The OAI-PMH annotation harvester was configured (using the Quartz scheduling library) to harvest updates to the annotation server every hour. The harvested annotation records are incorporated within the NLA metadata store, but saved as "annotation records", distinguishable from the original institutional metadata records. The link to the image is via the "target" field in the annotation metadata record. Table 1 below illustrates the original metadata record and an annotation record for an image in Picture Australia that is archived in the State Library of Victoria.

	NLA/State Library of Victoria Metadata	Annotation Record
Identifier	http://www.slv.vic.gov.au/pictoria/a23127.shtml	PILIN identifier
Title	House. Sydney. Harry Seidler. 1954-55	
Creator	Wille, Peter, photographer	Anna Gerber
Date	[ca. 1950-ca. 1973]	12 December 2007
Description	Colour slide of a Sydney house deigned by Harry Seidler	This was actually designed by Harry Seidler for his sister, Mary-Anne.
Subject	Slides	http://metadata.net/AustralianArchitecture.owl#Federation
Coverage	Sydney	
Rights	Reproduction rights owned by the State Library of Victoria	Creative Commons license
Source	State Library of Victoria	http://maenad:8080/Annotea/OAI1/
Type	image	annotation
Format	transparency : colour slide ; 35 mm	text
Target		http://www.slv.vic.gov.au/pictoria/a23127.shtml

Table 1: The Dublin Core Metadata Record and Annotation Record for a Sample Image

The web-based search and browse interface was extended to enable users to search across only authoritative metadata records, only annotations or both. Figure 4 illustrates the user interface that enables users to search for images using metadata fields from either the institutional metadata or annotation metadata schema. The annotation input, search and browse interface is a plug-in that is displayed on the left hand side of Figure 4. Users have the option to display a full-screen high resolution version of one of the retrieved result set – this also displays all of the attached annotations, with details of who created them and when (Figure 5).

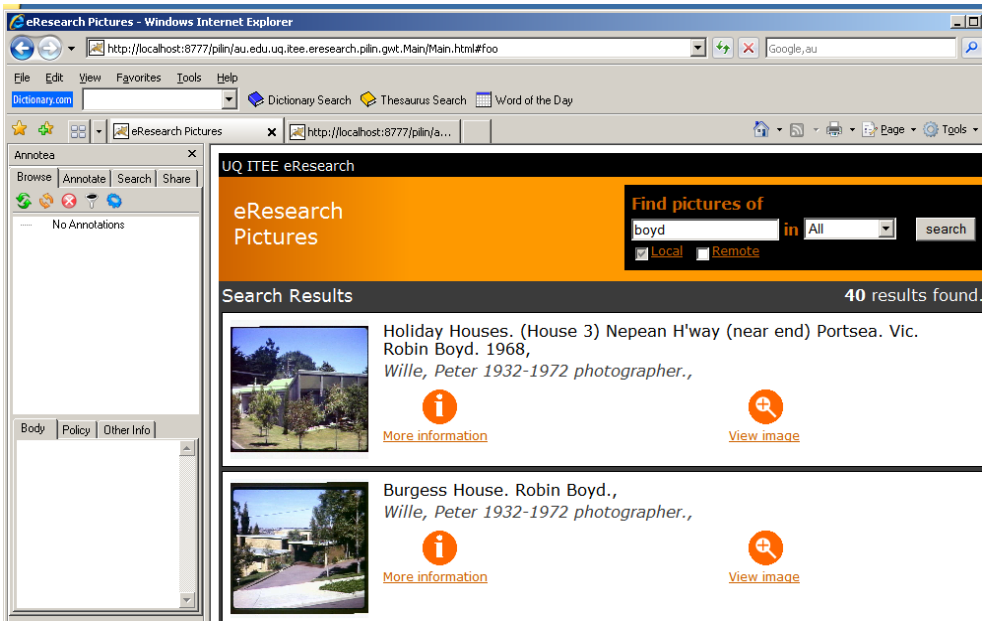


Figure 4: Screen Shot of the Enhanced PictureAustralia Search Interface

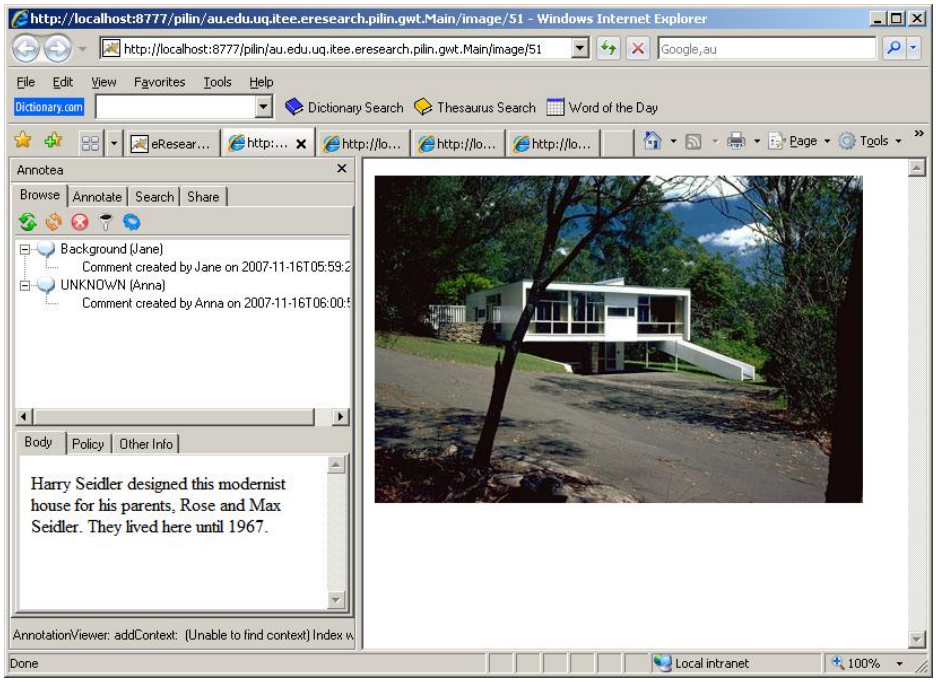


Figure 5: Screen Shot of the High Resolution Image and Annotation Browse Interface

Future Work and Conclusions

We are still in the process of evaluating and further developing HarvANA. We are currently working on: including tag clouds on the search interface; enabling the annotation of regions; semantic inferencing and querying based on the ontological tags; using social networks to rank query results. Feedback from potential user communities and repository service providers in response to demonstrations has been extremely positive. We are planning to carry out detailed user evaluations in collaboration with the UQ architecture department.

Through an OAI-PMH interface on Annotea, HarvANA delivers a scalable method by which custodians of collections or providers of federated search interfaces, can effectively and efficiently leverage community enthusiasm for collaborative social tagging systems. The adoption of an underlying schema and ontology-based tags that can be easily customized for specific communities, helps alleviate the problem of poor quality, inconsistent community-generated metadata. The security layer employs both authentication and access controls to help reduce the deliberate malicious attachment of offensive or erroneous tags. The resulting community-enhanced metadata and its representation in machine-processable RDF, enables more sophisticated and improved discovery and reasoning across existing open repositories.