

# HarVANA – Harvesting Community Tags to Enrich Collection Metadata

Jane Hunter, Imran Khan, Anna Gerber

University of Queensland

St Lucia, Queensland, Australia

(617) 33654311

{jane, imrank, agerber}@itee.uq.edu.au

## ABSTRACT

Collaborative, social tagging and annotation systems have exploded on the Internet as part of the Web 2.0 phenomenon. Systems such as Flickr, Del.icio.us, Technorati, Connotea and LibraryThing, provide a community-driven approach to classifying information and resources on the Web, so that they can be browsed, discovered and re-used. Although social tagging sites provide simple, user-relevant tags, there are issues associated with the quality of the metadata and the scalability compared with conventional indexing systems. In this paper we propose a hybrid approach that enables authoritative metadata generated by traditional cataloguing methods to be merged with community annotations and tags. The HarvANA (Harvesting and Aggregating Networked Annotations) system uses a standardized but extensible RDF model for representing the annotations/tags and OAI-PMH to harvest the annotations/tags from distributed community servers. The harvested annotations are aggregated with the authoritative metadata in a centralized metadata store. This streamlined, interoperable, scalable approach enables libraries, archives and repositories to leverage community enthusiasm for tagging and annotation, augment their metadata and enhance their discovery services. This paper describes the HarvANA system and its evaluation through a collaborative testbed with the National Library of Australia using architectural images from PictureAustralia.

## Categories and Subject Descriptors

H.3.5 [Online Information services]: Web-based services

H 3.1 [Content Analysis and Indexing]: Indexing methods

H 3.7 [Digital Libraries]: Dissemination, User issues

## General Terms

Performance, Design, Standardization

## Keywords

Social Tagging, Annotation, Harvesting, Metadata, Digital Collections, Ontology, Folksonomy

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '04, Month 1–2, 2004, City, State, Country.

Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

## 1. INTRODUCTION

Over the past few years, collaborative tagging and annotation systems that involve communities of users creating and sharing their own metadata, have exploded on the Internet. Sites such as Flickr [1], Del.icio.us [2], Connotea [3] and LibraryThing [4] are considered exemplary of the Web 2.0 phenomena [5] because they use the Internet to harness collective intelligence. Such systems provide a community-driven, “organic” approach to classifying information and resources on the Web, so that they can be browsed, discovered and re-used.

Proponents of social tagging systems [6-8] claim that because the terms used to describe the resources are community-defined, they are more topical, adaptive and relevant to users than traditional library cataloguing systems that use complex, relatively fixed, hierarchical thesauri and authority files. Terms in such controlled vocabularies do not evolve with popular language and many of them are irrelevant or anachronistic. Searches by non-experts often fail to yield results of relevance or that the users expect or understand. Authoritative metadata is also very expensive as it requires the time and effort of expert cataloguers. Social tagging and community annotation systems on the other hand, offer a mechanism by which the time consuming and expensive task of metadata generation can be distributed across communities. It is also argued that they provide a better measure of usefulness than software-based systems (e.g., Google) that rank resources based on the number of external links that point to a resource.

However, recent analyses of tagging data [9,10, 51], have shown that the indexing terms input by untrained users are often inconsistent and inaccurate – causing documents to go undiscovered or discovered in the wrong category. There is also the problem of scalability. Its difficult to predict how Flickr, del.icio.us, and other folksonomy-dependent sites will scale as content volume escalates. The flat structure of “folksonomies” [11] (long lists of simple tags) are useful for serendipitous browsing. But they don't support more sophisticated searching and browsing over very large collections. Folksonomies will not organically: evolve into synonymous clusters; identify preferred terms; or accrue into broader and narrower terms; features that are supported by thesauri and ontologies. Finally, a major limitation of social tagging systems is their lack of interoperability. Many of the popular social tagging systems are centralized, non-interoperable with other systems, don't support multiple levels of sharing and generally don't employ standards. Initiatives such as TagCommons [12] are investigating mechanisms and open standards to improve the interoperability of these popular community tools so tags can be shared across communities and resource types.

Despite these limitations, many cultural, scientific, and academic organizations responsible for providing access to large online collections recognize the potential value that community taggers can add to their collections. Projects such as the Commons [13] (a collaboration between Flickr and the Library of Congress) and Steve.Museum [14] are just two projects that aim to give online users a voice in describing the content of publicly-held collections, through online social tagging and annotation tools.

Our hypothesis (and that of the organizations involved in the Commons and Steve.Museum projects) is that significant value can be added to a collection by augmenting authoritative metadata with community-generated metadata that has undergone some form of quality control. However there are a number of unresolved challenges associated with implementing such a hybrid approach, that are deterring many organizations from incorporating community metadata in their metadata stores and search services. These include:

- How to carry out quality control of the community metadata without adversely impacting on the spontaneity and simplicity of the tags, and with minimal cost and effort?
- How to provide a measure of the authority of the source or author of the community-generated annotation? And how best to use this to rank retrieved results?
- How to access annotations and tags distributed over multiple different sites, systems and platforms – with different APIs?
- How to streamline the retrieval and aggregation of the community metadata with the authoritative metadata? There are no standards for representing or defining tags and annotations;
- How best to exploit, distinguish and display the community metadata through enhanced search, discovery and presentation services?
- How to identify terms from folksonomies that are candidates for inclusion in existing controlled vocabularies, to improve the relevance of authoritative metadata and the success rate of discovery services?

In this paper we describe the HarvANA (Harvesting and Aggregating Networked Annotations) system that we have developed at the University of Queensland. The objectives of HarvANA are to identify the optimum approach(es) for leveraging community annotation/tagging systems and to develop an efficient streamlined system (based on open standards and comprising a set of open source services) that enables collections managers to exploit the resulting metadata to improve discovery and reasoning across digital collections. In the process, the aim is to also identify solutions to the challenges listed above.

## 2. BACKGROUND

Web-based collaborative annotation systems are designed to enable online communities of users to attach descriptive terms to Web resources as a way of organizing the content for future navigation, filtering or search. When applied to digital resources shared via the Web, annotations provide a very powerful collaborative tool - enabling the capture and dissemination of individuals' and group opinions about particular digital resources.

There are a huge variety of both annotation types and systems for creating and re-using them. A number of researchers have attempted to identify dimensions by which to categorize both annotations [15,16] and the plethora of annotation systems [17-

19]. Annotation systems vary from: stand-alone applications for personal collections [20,21]; to extensions to proprietary systems [22]; to digital library-focussed systems [23] and general Web-based systems [24]. In the context of the work described here we are interested in collaborative web-based annotation and tagging systems. But even within this scope, there are a huge number of systems, for annotating a wide range of target resources.

A significant number of annotation systems have been designed for annotating specific media or document types (e.g., Web pages, images, video, Wikis) and include specific client functionality optimized for the particular media type. Some systems only allow annotations to be attached to whole files or specific types of segments (e.g., keyframes) whilst other systems provide tools that enable users to interactively specify the segments or regions or components to be annotated. Most systems provide client interfaces that allow users to create, edit, delete and query annotations, as well as view and browse existing annotations. Some systems allow users to respond to existing annotations generating hierarchical annotation threads. Many systems enable users to be automatically notified of new annotations on particular topics or by particular users by subscribing to RSS/Atom feeds.

### 2.1 Informal versus Formal Annotations

Annotations may be informal (free text) or formal. Formally defined annotations conform with a hierarchical set of fields and values specified within an XML Schema or RDF Schema and controlled vocabularies. Typically an annotation record contains: a body (e.g., one or more keywords; and/or a free text description); the name of the author; date of attachment; and a link to the annotated resource or a pointer to a location or anchor within the resource.

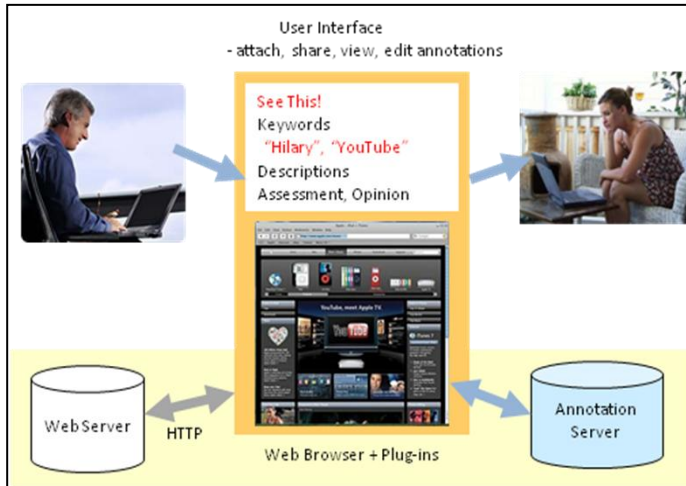
In some cases, the annotations are also “ontologically” defined – the structure and formally defined constituents are based on concepts defined within a given OWL ontology [25]. Ontology-based annotations are valuable because, in addition to validation and quality control, they allow reasoning about the object they are annotating. Within the context of this paper we refer to annotations represented using RDF [26] or OWL as “semantic annotations” – because this mark-up enables the annotated resources to become accessible to the larger Semantic Web.

### 2.2 Tags versus Annotations

“Tags” are an important subclass of annotations that comprise simple, unstructured labels or keywords assigned to digital resources to describe and classify the digital resource (e.g., a web page, image or blog post). Tags allow the user to organize resources into categories (groups of resources with the same tag) so they can be more easily retrieved later. Users can define their own tags based on the terms that seem most relevant at the time and can choose to share them with others in their social network. The tags produced by a community of users evolve organically over time and can be aggregated to form a community vocabulary, christened a “folksonomy” by Vander Wal [27,28]. Such social tagging approaches have also been described as ‘ethnoclassification’ [29] or ‘distributed classification’ systems [30]. Hammond et al [31] describe social tagging as a ‘bottom-up’ or grass-roots approach compared with the traditional top-down approach of institutional classification based on formal and less flexible classification systems.

## 2.3 Architectural Models

Most Web annotation systems comprise: an annotation creation/authoring and attachment interface; an annotation browse, search and retrieval interface; and an annotation storage and indexing component. Figure 1 provides a high-level view of a collaborative Web annotation system.



**Figure 1: High level View of a Web Annotation System**

The main dichotomy between system architectures, is whether the annotations are stored separately from the documents they are annotating – or whether they are stored together on the same centralized site. Flickr is an example of a centralized application that stores the resources and the tags on the same system and that requires both resource providers and users to submit their contributions via the one site. The separate storage approach calls for Web servers and data storage dedicated to annotations which include hyperlinks to the original target resources (via URIs). Although separate storage carries overheads associated with maintaining HTTP access to indexed storage repositories and web servers, and maintaining links between annotations and documents, it has a number of advantages. The decoupling of the annotations from the content allows more control and flexibility over how the annotations are accessed, processed, presented and re-used. If the annotations are on a separate server, then access, authorization and posting of responses can more easily be controlled and restricted to a particular community of users. Separation allows a single resource to be annotated in many different ways by users on the same annotation server or on different annotation servers, using different community-specific terminologies or ontologies. Separating the annotations from the resources also avoids the copyright issues that arise when having to store a copy of the digital resources on the social tagging site.

## 3. RELATED WORK

### 3.1 Metadata Aggregation

The importance of being able to aggregate distributed metadata from a range of sources has been recognized by a number of projects. Annotea [39], TagCommons [12], Steve.museum [14] and Rich Tags [58] have all recognized the need for standardized ways of defining annotations and tags so they can be shared between communities. However the problem of annotation aggregation is still largely unresolved. The challenge is that a

central server has to download remote RDF annotations to the local server as quickly and efficiently as possible, whilst maintaining synchronization with the most recently added remote annotations and performing data merging, inferencing and querying over the integrated data sets – perhaps with the inclusion of some ranking or weighting mechanism to give more or less weight to those annotations provided by most trusted colleagues or organizations. Previous approaches to querying and aggregating distributed (RDF) annotation servers can be divided into two categories:

- RSS Feeds – using this approach, the central agency subscribes to RSS feeds [32] from the distributed annotation servers. Because this approach involves the continual transmission of small amounts of RDF data associated with any updates, it is not controllable by the agency receiving the feeds and is not scalable when databases of considerable dimensions are involved.
- Peer-to-peer approaches - each annotation server is considered as a peer with its own local RDF database. RDF queries are broadcast to each peer and the results are aggregated on retrieval. Systems like Edutella [33] and the RDFPeers framework [34] suffered from scalability issues because of the computational burden when many users are simultaneously querying each server. However, there are a couple of projects [35,36] that are investigating specific algorithms to improve the scalability of annotation exchange in peer-to-peer applications.

Within the HarvANA project, we propose a novel, third approach - retrieving annotation data stored on multiple distributed servers via an OAI-PMH [37] interface sitting on top of the annotation servers. This approach involves mapping the annotations stored on an Annotea server to the Dublin Core metadata schema and periodically harvesting them by having the central agency send OAI-PMH (HTTP) requests to the server. The advantages of this approach include:

- It is controlled by the harvesting organization;
- It can be configured to run periodically at set times (e.g., hourly or nightly when network traffic is low);
- Date stamps (*from* and *until*) can be used to only harvest updates since the last harvest;
- It provides a platform-independent, standardized approach that is independent of the actual tagging system or the type of resources being tagged;
- OAI-PMH has been widely adopted and has been shown to be scalable for large distributed collections [38].

### 3.2 Quality Control

Despite the enthusiasm and hype around social tagging systems, recent analyses [9,10] have revealed that there is a significant degree of idiosyncrasy, inconsistency, contradiction and inaccuracy within folksonomic tags. Significant numbers of tags include: misspellings, poor encoding, acronyms, punctuation and compound tags that omit spaces. Both surveys also found a high proportion of tags labelled “toread”, “todo”, “fun” and “cool”. Such erroneous and non-descriptive tagging may not matter if the number of taggers reaches “critical mass” but in many systems, the number of tags may be very low and such errors will make resources difficult to find. A second major disadvantage of distributed classification systems is their flat structure and consequent inability to handle synonyms (multiple tags with the

same or similar meaning) and homonyms (a single tag with multiple meanings). The lack of hierarchical structure (sub-class relationships) between tags fails to identify relationships between resources of the same or similar type, again leading to reduced precision and recall.

Significant research effort is currently focused on mechanisms by which tags can be improved and more useful formal semantics can be derived from simple community tagging systems. The aim is to optimize the trade-off between the simplicity and freedom of community tagging and the benefits to search engines of hierarchical structured vocabularies. Suggested approaches have included supervised folksonomies – in which taxonomists work with an underlying folksonomy that has evolved through community participation to add syntax and structure [52]. Another approach is “ontology-directed-folksonomy” in which users are provided with suggested and popular tags from an ontology, but still have the option to define their own tags [53-55]. Guy and Tonkin [9] and Mejias [56] suggest a combination of post-processing to “clean-up” the tags together with the adoption of best practise guidelines. The paradox is that such guidelines will impact adversely on the freedom and ease of use of the system. Another approach to managing the large numbers of tags that accumulate over time and usage is to cluster or group them – either *a priori* by the users or *a posteriori* using clustering algorithms. Suggestica’s RawSugar [57] is one example of the *a priori* approach.

For the HarvANA system we decided to adopt the “ontology-directed-folksonomy” approach. When entering tags, users are provided with suggested and popular tags from an ontology (specified at system configuration) but they still have the option to define their own unique tags. This approach is stable, flexible, ensures maximum semantic richness of the metadata and facilitates easy adaptation to a different community simply by changing the backend ontology. A pull-down menu provides the interface to the ontology when inputting and searching on tags. When a user searches on a parent tag, all items with the parent tag, synonym tags or children tags are retrieved. The class hierarchies are also incorporated within the tag cloud to embed multi-level structuring.

In addition, HarvANA restricts access to the annotation server via Shibboleth identity management. Our approach is to provide annotation services for a closed community with specific knowledge or expertise - rather than the general public. This reduces the proportion of incorrect, inappropriate or misleading tags and obviates the need for a moderator to check annotations.

## 4. OBJECTIVES

Discussions with the National Library of Australia (who were keen to be involved in a pilot project investigating the value of community tagging to public collections) led to the proposal of a collaborative project with the following objectives:

- To identify a common model for modeling tags and annotations across Web-based collaborative systems
- To develop a standardized approach for retrieving, aggregating, using and presenting metadata attached by a number of distributed sources e.g., authoritative institutional metadata as well as annotations and tags attached by different users using the same or different systems
- To develop and evaluate an architecture that involves adding an OAI-PMH layer to an annotation server to periodically

harvest community, distributed metadata – social tags and annotations

- To develop a set of easily deployable tools and services for attaching annotations, harvesting annotations, aggregating distributed annotations with metadata, searching and browsing aggregated metadata and presenting search results and associated annotations
- To investigate the optimum approach to making both the authoritative metadata and community tags/annotations accessible to the search engine to enhance search services
- To investigate the optimum user interface for presenting (but distinguishing between) both the metadata and annotations/tags associated with retrieved resources
- To evaluate the system through extensions to: an existing community annotation system (Co-Annotea) [44]; a testbed collection of images from the PictureAustralia collection hosted by the National Library of Australia (NLA)[48]; and the participation of and feedback from a particular community of expertise.

The remainder of this paper describes the outcomes of the project within the scope of these objectives.

## 5. A COMMON MODEL

A common extensible model for representing annotations and tags is essential to ensure compatibility and interoperability of annotation and tagging systems and the sharing and re-use of tags generated within different communities and tagging sites. Although a number of initiatives have proposed alternative models, the W3C’s Annotea RDF model [39] is emerging as a defacto standard for modelling annotations and tags – having been adopted by a large array of both clients and servers, including: Annotea, which is an Annotea implementation for the Mozilla browser [41]; annoChump [42]; Vannotea, an Annotea-based collaborative annotation system for multimedia objects [43]; and Zannot, a Zope annotation server based on Annotea [40].

The Annotea model specifies the following attributes associated with an annotation:

- *body* (the actual textual description or tag value(s));
- *type* (the top-level class is *annotation*, but possible sub-classes include: *comment*, *query*, *review*, *rating*, *assessment*);
- *creator* (the author of the annotation or tag);
- *date\_created* (the date the tag/annotation was attached and published).

Extensions are possible through the addition of further optional attributes that might include fields such as: *language*, *media\_type*, *format* (e.g., the annotation may be a URL, audio, image, video). XPointer [45] provides a standardized method for locating annotations within XML and HTML documents but is inadequate for fine grained specification of regions or segments within multimedia resources e.g., the region of an image, a scene from a video. Such segments require media-dependent locators that are specifically defined through a new attribute e.g., the “extent” attribute in Figure 3. In addition, security in the form of access control can be implemented by the addition of an “access\_policy” attribute, which points to a machine-interpretable XACML policy [46] associated with the annotation itself. Such a policy specifies who can read, edit or delete the annotation. The enforcement of such a policy relies on the authentication and identification of users of the system through an Identity Management system such as Shibboleth [47].

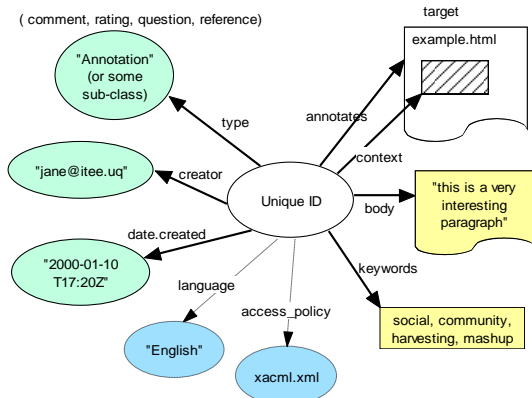


Figure 2: Annotea RDF Model for Web Annotations

## 6. SYSTEM ARCHITECTURE

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, to enrich the metadata store with community knowledge. Figure 3 provides a high-level view of the HarvANA system architecture.

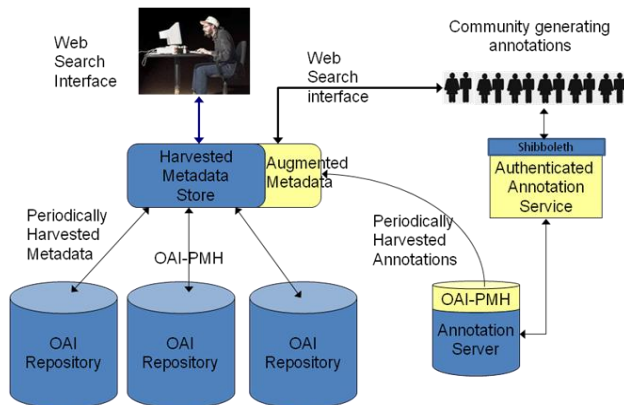


Figure 3: High-level View of the HarvANA Architecture

In addition, a security interface based on Shibboleth user authentication [47] and XACML access controls [46] has been implemented which restricts access to the annotation server (and individual annotations) to the members of specific online communities [57].

### 6.1 Underlying Technologies

HarvANA uses W3C's Annotea annotation protocol [39] and an RDF Jena data store for storing and querying annotations. 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. Users can search across annotation attributes that include: creator, date, keywords or free-text searching over the description. Quality control of the annotations is provided by validating annotations/tags against the schema and restricting tags to the specified 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 are then incorporated within the original institutional metadata store and indexed using Lucene.

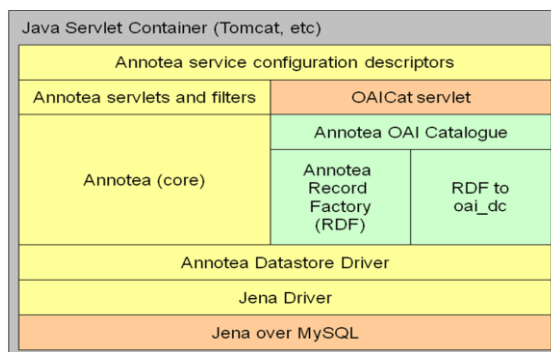


Figure 4: Technical Components of HarvANA

## 7. TESTBED AND CASE STUDY

We are currently evaluating HarvANA through a collaboration with the National Library of Australia (NLA). Our aim is 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. The aim is to provide annotation services to a community of experts with specific knowledge that would enhance the descriptive metadata for a specific collection.

PictureAustralia is an NLA project that provides a federated discovery service to more than 1 million images from 31 contributing organizations [48]. 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 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 [58] that limits the keywords or tags to a set of controlled, machine-processable terms (Figure 9). Our "community of experts" were colleagues from the architecture department at the University of Queensland. 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 “relation” field in the annotation DC metadata record. Table 1 shows both an authoritative metadata record and an annotation record for an image accessible via PictureAustralia but held in the State Library of Victoria.

## 7.1 User Interface

### 7.1.1 Creating and Editing Annotations

Figure 5 illustrates the Annotea sidebar ( a client plug-in for IE and Firefox) that provides a browser-based interface for creating and attaching annotations to images retrieved via the NLA’s PictureAustralia Web search interface. Users specify a *title*, *type* and the actual content (free text description and/or keywords).

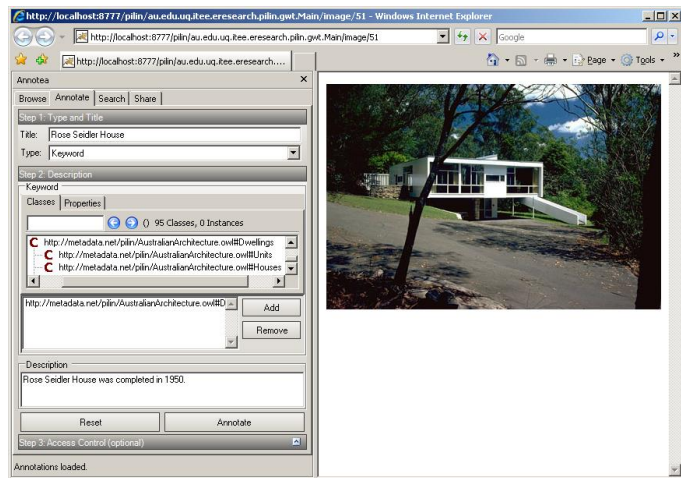


Figure 5: Creating an Annotation using the Annotea sidebar

Keywords or tags can either be selected/auto-completed from a pre-defined ontology (of architectural terms) or newly defined. When creating a new annotation, the user may also define an (XACML) access policy or attach a Creative Commons license. The newly created annotation is validated against the RDF model and assigned a unique persistent identifier, before being saved.

### 7.1.2 The Enhanced Search & Presentation Interface

The NLA’s existing web-based search interface to the PictureAustralia image collection was extended to enable users to search across only institutional metadata, only community annotations or both. Figure 6 illustrates the extended search interface. If the option to search community annotations is selected, users may choose to search on *tags*, *description*, *subject*, *title*, *date* or *creator* fields. For example, a user may search for all resources annotated by 'Anna Gerber'. Selecting the “more information” button beside each thumbnail displays the authoritative metadata above the annotations (in chronological order) using colours to distinguish between them (Figure 7).

### 7.1.3 Tag Cloud Browsing

A tag cloud showing the most popular tags is generated and displayed at the top of the search page (Figure 8). Clicking on a tag in the cloud triggers a search for items with that tag, synonyms or children tags. When the results are displayed, a related tag cloud (showing tags that are related by subClass or equivalentClass relationships) is displayed at the bottom of the

page. The search can be refined by clicking on a related tag name, or expanded to search on all related tags.



Figure 6: Enhanced Picture Australia Search Interface

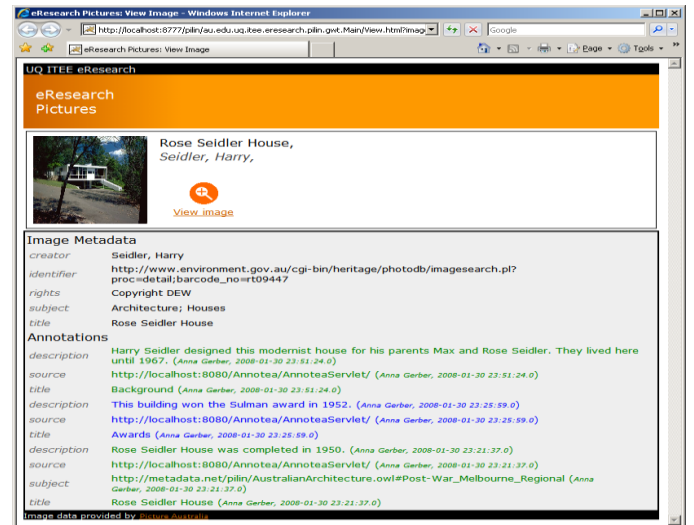


Figure 7: Displaying Library and Community Metadata

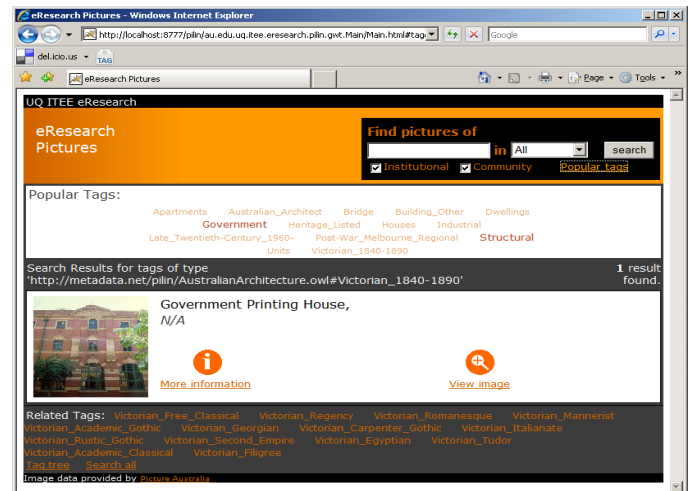


Figure 8: Browsing via Popular and Related Tag Clouds

## 8. EVALUATION

The aim of HarvANA is to enhance the discovery services for a large online collection, by harvesting tags and annotations attached by a specific community, and aggregating them with the existing authoritative or institutional metadata to enrich it and improve findability. In order to evaluate the effectiveness of the approach, we evaluated three key criteria:

- Usability of the system from an end-user point of view;
- Improvements to the search and discovery services;
- System design, efficiency and deployment from an administrative point of view.

End-users of the system comprised two types: the authenticated community members who attach the annotations and the general public who search the collection. Users attaching the annotations found the Annotea browser plug-in, easy to download, install and configure but requested auto-completion of the ontology-based tags as well as via a pull-down multi-level menu. They also requested the ability to see existing metadata and tags for the current resource they are annotating. Currently this is only visible whilst searching. They also requested the ability to annotate regions of images. Feedback from users searching the collection, indicated that addition of the tag cloud to the PictureAustralia search interface was a positive improvement. They felt it added a sense of social interaction, non-conformism and ‘fun’ to the interface. Although, many users said that they would continue to use the free text search field. Feedback on relevance of search results indicated that the addition of ontology-based tags by domain experts, improved the discovery of relevant images, for both naïve and expert users, as well as enriching the collection’s inherent educational value. Feedback from a collection management perspective was provided by the NLA. Collections managers who were surveyed, approved of the use of OAI-PMH and Dublin Core, with which they are already familiar. The use of OpenID instead of Shibboleth for the user authentication was raised. They believe there needs to be more detailed evaluation with respect to the value-add of the community annotations. The separation of the annotation servers from the centralized metadata store was seen as an advantage as it removed them of responsibility for hosting and maintaining the annotation servers. The ability to easily reconfigure the system and add new annotation servers to be harvested was seen as a key benefit.

## 9. FUTURE WORK

To date, we have implemented and evaluated what we consider to be a beta version of HarvANA. This process has revealed a number of key areas that warrant further investigation and development:

- HarvANA has been designed to support the annotation of images within PictureAustralia – we are interested in how the system would need to be modified to support the annotation of other document types (e.g., MusicAustralia) and parts of documents e.g., regions of images;
- The incorporation of social network information - FOAF (Friend-Of-A-Friend) trust profiles [49] enable users to rate how much they trust their web of acquaintances on particular subjects. The trust ratings can be used to rank search results based on tags by trusted colleagues. For example, within FilmTrust, Golbeck et al [50] use social networks to

personalize search results for films based on recommendations by trusted reviewers;

- Identifying tags that should be incorporated in the ontology. Collections managers need tools to assist them to identify which tags to incorporate and where they should be added;
- Investigating the use of SPARQL to perform semantic inferencing and querying and to derive new knowledge;
- The harvested annotations are of significant interest to the regional libraries who actually host the PictureAustralia images. We are interested in using RSS feeds to stream the harvested annotations back out to these and other sites;
- Machine-learning - in recent years a number of systems have combined machine-learning techniques with manual ontology-based annotation to index web pages, textual documents and multimedia. However they fail to incorporate or leverage community-driven social tagging systems. We are keen to investigate tripartite classification systems that combine social tagging, machine-learning and traditional library classification approaches, integrated through common structured ontologies.

Finally, the work to date, has focused on the annotation of images from a public library collection by a closed community of experts. There are many scenarios in which this approach would be of value and which would undoubtedly reveal a different set of issues and problems. For example, we are keen to explore HarvANA’s application to: the annotation of museum collections by indigenous communities; peer-review of scholarly publications by particular disciplinary experts; the annotation of medical images by domain experts (e.g., teledermatology).

## 10. CONCLUSIONS

In this paper we have proposed, developed and evaluated a standards-based, platform-independent approach by which social tagging and annotation systems can expose their tag data so it can be readily accessed, harvested and re-used by meta-search services. By opening up tagging data through a combination of: a standardized model, a harvesting protocol and a metadata mapping, we enable both the custodians and users of digital repositories to benefit from the enormous potential value of collaborative tagging, with a minimum of effort and no prior knowledge of the backend annotation systems. The advantages of our approach include:

- The separation of annotation servers from the central metadata store and actual digital objects, enables the easy incorporation of existing or new annotation and tagging sites that can be managed by communities external to the library, museum or archive;
- Because the annotations are stored separately to the digital objects they describe (but point to via URIs), we do not have to be concerned with copyright issues that arise when the digital objects have to be copied to the social tagging site (e.g., Flickr and the Library of Congress Commons Project)
- The common RDF model for representing annotations/tags together with OAI-PMH:
  - provides a standardized mechanism by which sites can expose their tag data, so that it can be harvested and re-used by metasearch services.
  - enables machine-understanding of the tags and annotations and more sophisticated semantic inferencing by meta-search services;

- enables aggregation of tags from multiple distributed annotation servers, regardless of the underlying tagging platform or system.
- The decision to use the already widely adopted OAI-PMH and Annotea protocols provides a high level of openness, interoperability and low barrier to entry;
- The adoption of Shibboleth and XACML provide a means of authenticating the source of the annotations and restricting access to, editing and re-use of the annotations based on author-specified policies. These security mechanisms also help protect against malicious taggers and tag spam;
- The “ontology-directed-folksonomy” approach (using a community-defined light-weight ontology to suggest keywords) provides quality control at time of capture, and reduces the need for post-processing without restricting the freedom of taggers;
- The flexible system design enables easy re-configuration and customization. Administrators can easily modify the OWL ontology that defines community-specific tags or the frequency of harvesting of the annotations. They can also easily add new annotation servers to be harvested.

By combining OAI-PMH with Annotea, HarvANA delivers a scalable, interoperable method by which custodians of collections can effectively and efficiently leverage community enthusiasm for collaborative tagging systems. The adoption of a common model for representing annotations and a light-weight community-specific ontology for tags provides the optimum combination of formal and informal metadata and helps to address the current weaknesses in tagging systems. By combining the best aspects of the Social and Semantic Webs, HarvANA represents an early exemplar of the Web 3.0 paradigm – a collective knowledge system [62].

## 11. ACKNOWLEDGMENTS

This work was funded by DEST through the Systemic Infrastructure Initiative (SII) ARCHER and PILIN projects. Our thanks also to the National Library of Australia (NLA) for providing us with access to images and metadata from the PictureAustralia collection and for their valuable feedback.

## 12. REFERENCES

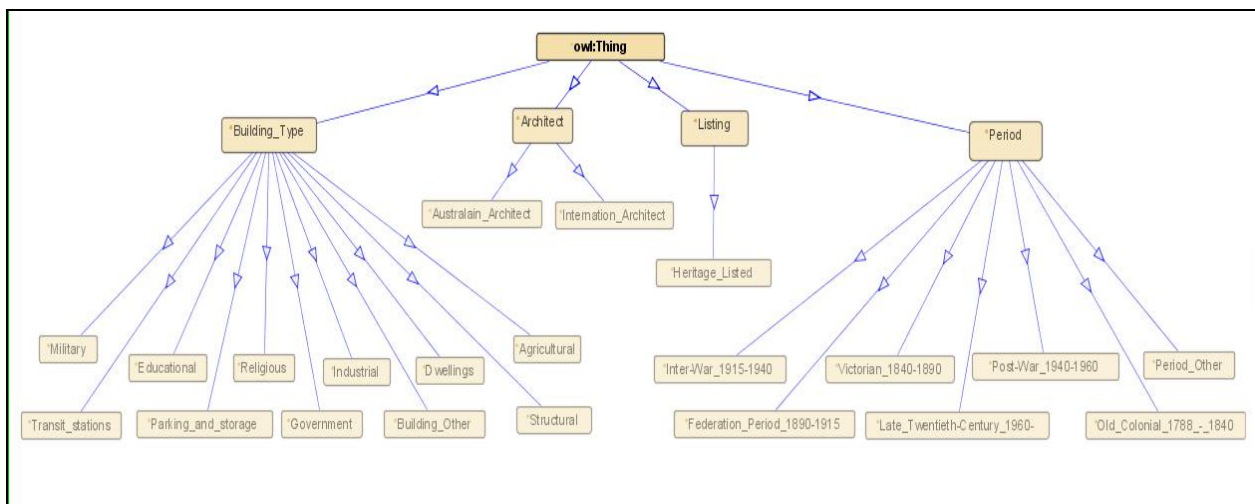
- [1] Yahoo! Inc. 2008. Flickr <http://www.flickr.com/>
- [2] Del.icio.us (2008). Del.icio.us: Social Bookmarking. <http://del.icio.us/>
- [3] Lund, B., Hammond, T., Flack, M., and Hannay, T. 2005. Social Bookmarking Tools (II): A Case Study – *Connotea*, *D-Lib Magazine* **11**(4), April 2005. <http://www.dlib.org/dlib/april05/lund/04lund.html>
- [4] LibraryThing. 2008. <http://www.librarything.com/>
- [5] O'Reilly T. 2005. What Is Web 2.0. O'Reilly Network. September 30, 2005 <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
- [6] Shirky, C. 2005. *Ontology is Overrated: Categories, Links, and Tags*. Retrieved 17 January, 2007, from [http://www.shirky.com/writings/ontology\\_overrated.html](http://www.shirky.com/writings/ontology_overrated.html)
- [7] Merholz, P. 2004. *Metadata for the masses*. Retrieved 17 January, 2007, from <http://www.adaptivepath.com/publications/essays/archives/000361.php>
- [8] Kroski, E. 2005. *The Hive Mind: Folksonomies and user-based tagging*. Retrieved 17 January, 2007, from <http://infotangle.blogspot.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/>
- [9] Guy, M., and Tonkin, E. 2006. “Folksonomies: Tidying up tags?”, *D-Lib Magazine*, Volume 12, Number 1, January, 2006 <http://www.dlib.org/dlib/january06/guy/01guy.html>
- [10] Kipp, M. E. and Campbell, G. D. 2006. 'Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices', Proceedings of the Annual General Meeting of the American Society for Information Science and Technology, Austin, TX, November 3-8, 2006.
- [11] Vander Wal, T. 2005. Explaining and showing broad and narrow folksonomies. Feb 21, 2005. [http://personalinfocloud.com/2005/02/explaining\\_and\\_.html](http://personalinfocloud.com/2005/02/explaining_and_.html)
- [12] TagCommons 2007. Ontologies Vs. Formats Vs. Schema Vs. API. March 2, 2007. <http://tagcommons.org/>
- [13] The Commons 2008 – The Flickr and Library of Congress Pilot Project <http://www.flickr.com/commons>
- [14] Chun, S, Cherry R, Hiwiller D, Trant J. and Wyman, B 2006. “Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums,” *Museums and the Web* 2006.
- [15] Marshall, C. 1998. [Toward an ecology of hypertext annotation](#) in *Proceedings of ACM Hypertext '98*, Pittsburgh, PA (June 20-24, 1998) pp. 40-49.
- [16] Reeve, L. and Han, H. 2005. Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM Symposium on Applied Computing* (Santa Fe, New Mexico, March 13 - 17, 2005). L. M. Liebrock, Ed. SAC '05. ACM, New York, NY, 1634-1638. DOI=<http://doi.acm.org/10.1145/1066677.1067049>
- [17] Sazedj P. and Pinto, H.S. 2005. Time to evaluate: Targeting Annotation Tools, *Semannot 2005*, Nov. 2005 .
- [18] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. Ciravegna, F., 2006. 'Semantic annotation for knowledge management: Requirements and a survey of the state of the art', *Web Semantics: Science, Services and Agents on the World Wide Web* , vol. 4, no. 1, 14-28 (2006).
- [19] Speller, E., Collaborative tagging, folksonomies, distributed classification or ethnoclassification: a literature review [http://informatics.buffalo.edu/org/ljsj/articles/speller\\_2007\\_2\\_collaborative.php](http://informatics.buffalo.edu/org/ljsj/articles/speller_2007_2_collaborative.php)
- [20] Chen, C., Oakes, M., and Tait, J. 2006. A location annotation system for personal photos. In *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Seattle, Washington, USA, August 06 - 11, 2006). SIGIR '06. ACM, New York, NY, 726-726. DOI: <http://doi.acm.org/10.1145/1148170.1148339>
- [21] Fu, X., Ciszek, T., Marchionini, G. and Solomon, P. 2005. Annotating the Web: An Exploratory Study of Web Users'

- Needs for Personal Annotation Tools. In Grove, Andrew, Eds. *Proceedings 68th Annual Meeting of the American Society for Information Science and Technology (ASIST)* 42, Charlotte (US).
- [22] Groza, T., Handschuh, S., Möller K. and Decker, S. 2007. SALT - Semantically Annotated LaTeX for scientific publications. In *Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*. Innsbruck, Austria, 2007
- [23] Agosti, M. Ferro, N., Frommholz, I., Panizzi, E., Putz, W. and Thiel, U. 2006. Integration of the DiLAS Annotation Service into Digital Library Infrastructures. In: *Proc. of the Workshop on Digital Libraries in the Context of Users' Broader Activities (DL-CUBA 2006)*. June 2006, Chapel Hill, NC, USA.
- [24] Fernandes, M., Alho, M., Martins, J. A., Pinto, J. S., and Almeida, P. 2005. Web Annotation System Based on Web Services. In *Proceedings of the international Conference on Next Generation Web Services Practices* (August 22 - 26, 2005). NWESP. IEEE Computer Society, Washington, DC.
- [25] W3C, 2004. OWL Web Ontology Language Overview, W3C Recommendation, Eds. D.McGuinness, F. van Harmelen, 10 February, 2004 <http://www.w3.org/TR/owl-features/>
- [26] W3C, 2004. Resource Description Framework (RDF). RDF Core Working Group, 2004. <http://www.w3.org/RDF/>
- [27] Porter, J. 2005. *I've heard of folksonomies, Now how do I apply them to my site?* Retrieved 17 January, 2007, from [http://www.bokardo.com/archives/applying\\_folksonomies/](http://www.bokardo.com/archives/applying_folksonomies/)
- [28] Smith, G. 2004. *Folksonomy: social classification*. Retrieved 17 January, 2007, from [http://atomiq.org/archives/2004/08/folksonomy\\_social\\_classification.html](http://atomiq.org/archives/2004/08/folksonomy_social_classification.html)
- [29] Merholz, P. 2004. *Ethnoclassification and vernacular vocabularies*. Retrieved 17 January, 2007, from <http://www.peterme.com/archives/000387.html>
- [30] Mejias, U. A. 2004. *Bookmark, classify and share: A mini-ethnography of social practices in a distributed classification community*. Retrieved 17 January, 2007, from [http://ideant.typepad.com/ideant/2004/12/a\\_delicious\\_stu.html](http://ideant.typepad.com/ideant/2004/12/a_delicious_stu.html)
- [31] Hammond, T., Hannay, T. Lund, B., Scott, J., 2005. Social Bookmarking Tools A General Review , *D-Lib Magazine*, April 2005, Volume 11 Number 4. <doi:10.1045/april2005-hammond>.
- [32] Winer, D. 2007. RSS 2.0 at Harvard Law. RSS 2.0 Specification. <http://cyber.law.harvard.edu/rss/rss.html>
- [33] Nejdil, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., and Risch, T. 2002. EDUTELLA: a P2P networking infrastructure based on RDF. In *Proceedings of the 11th international Conference on World Wide Web* (Honolulu, Hawaii, USA, May 07
- [34] Cai M. and Frank, M. 2004. RDFPeers: A Scalable Distributed RDF Repository based on A Structured Peer-to-Peer Network. In *International World Wide Web Conference (WWW)*, 2004. <http://citeseer.ist.psu.edu/cai04rdfpeers.html>
- [35] Tummarello, G. Morbidoni, C. Bachmann-Gmür, R. Erling, O. 2007. "RDFSyc: efficient remote synchronization of RDF models", ISWC 2007, Korea, November 2007
- [36] Heine, F. 2006. "Scalable P2P based RDF Querying", ACM International Conference Proceeding Series; Vol. 152 , Proceedings of the 1st international conference on Scalable information systems, Hong Kong Article No. 17 Year of Publication: 2006 ISBN:1-59593-428-6
- [37] Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. "The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0". June 2002. <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>
- [38] Henry, J., Liu, X., Hochstenbach, P. and Van de Sompel, H. 2004. "The multi-faceted use of the OAI-PMH in the LANL Repository," *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, June 7-11 2004, Tuscon, AZ, USA*. pp 11-20.
- [39] Koivunen, M.-R. and Kahan, J. 2001. Annotea: an open RDF infrastructure for shared Web annotations. In *Proceedings of the 10th international conference on World Wide Web*. Hong Kong. ACM Press (2001)
- [40] Rhaptos, 2004. Zannot: Zope Annotea Server. <http://rhaptos.org/downloads/zope/zannot/>
- [41] Mozdev, 2008. Annozilla (Annotea on Mozilla). <http://annozilla.mozdev.org/>
- [42] W3C 2001. annoChump Overview. 5 December, 2001. <http://www.w3.org/2001/09/chump/>
- [43] Schroeter, R., Hunter, J., and Kosovic, D. 2003. Vannotea - A Collaborative Video Indexing , Annotation and Discussion System For Broadband Networks. In *Knowledge Markup and Semantic Annotation Workshop, K-CAP 2003*. Sanibel, Florida (2003)
- [44] Schroeter, R., Hunter, J., Guerin, J., Khan I. and Henderson, M. 2006. "A Synchronous Multimedia Annotation System for Secure Collaboratories" *2nd IEEE International Conference on E-Science and Grid Computing (eScience 2006)*. Amsterdam, Netherlands. December 2006. p 41.
- [45] W3C 2002. XML Pointer Language (XPointer). W3C Working Draft 16 August, 2002. <http://www.w3.org/TR/xptr/>
- [46] Oasis 2008. eXtensible Access Control Markup Language (XACML), [http://www.oasis-open.org/committees/tc\\_home.php?wg\\_abbrev=xacml](http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml)
- [47] Internet2 2008. Shibboleth. <http://shibboleth.internet2.edu/>
- [48] National Library of Australia PictureAustralia
- [49] Brickley D. and Miller L. 2007. FOAF Vocabulary Specification 0.91. November 2007 <http://xmlns.com/foaf/spec/>
- [50] Golbeck J. 2006. Generating Predictive Movie Recommendations from Trust in Social Networks. *Proceedings 4th International Conference, iTrust 2006*, Pisa, Italy. pp 93-104
- [51] Golder, S. and Huberman B.A. 2006. "Usage Patterns of Collaborative Tagging Systems." *Journal of Information Science*, 32(2). 198-208

- [52] Rosenfeld, L. 2005. Folksonomies? How about Metadata Ecologies?  
[http://louisrosenfeld.com/home/bloug\\_archive/000330.html](http://louisrosenfeld.com/home/bloug_archive/000330.html)
- [53] Pind, L. 2005. *Folksonomies: How we can improve the tags*. Retrieved 17 January, 2007, from  
<http://pinds.com/articles/2005/01/23/folksonomies-how-we-can-improve-the-tags>
- [54] Vuorikari, Riina 2007. Folksonomies, social bookmarking and tagging: the state-of-the-art, Special Insight Reports, [http://insight.eun.org/shared/data/insight/documents/specialreports/Special\\_Report\\_Folksonomies.pdf](http://insight.eun.org/shared/data/insight/documents/specialreports/Special_Report_Folksonomies.pdf)
- [55] Microsoft Research 2008. TagBooster: A System for Ranking and Suggesting tags.  
<http://research.microsoft.com/~milanv/tagbooster.htm>
- [56] Mejias, U. A. 2005. *Tag literacy*. Retrieved 17 January, 2007, <http://blog.ulisesmejias.com/2005/04/26/tag-literacy/>
- [57] Khan, I., Schroeter R. and Hunter, J. 2006. "Implementing a Secure Annotation Service", *International Provenance and Annotation Workshop*, Chicago, USA. 3 - 5 May 2006.
- [58] Smith, D. A., Lambert, J. and schraefel, m. c. 2008. *Rich Tags: Cross-Repository Browsing*. In: Open Repositories Conference 2008 (OR 2008), April 2008, Southampton, UK
- [59] Apperly R., Irving R. and Reynolds P. 1989. A pictorial guide to identifying Australian architecture : styles and terms from 1788 to the present. Angus & Robertson, 1989.
- [60] Hearst M. and Rosner, D. 2008. *Tagclouds: Data Analysis tool or Social Signaller?* HICSS 2008, Social Spaces Minitrack, January 2008, Hawaii  
<http://flamenco.berkeley.edu/papers/tagclouds.pdf>
- [61] HarvANA Demo <http://maenad.itee.uq.edu.au:8080/harvana/>
- [62] Gruber T. 2007. Collective Knowledge Systems: Where the Social Web meets the Semantic Web. *Journal of Web Semantics* (2007), doi:10.1016/j.websem.2007.11.011

**Table 1: The Dublin Core Metadata Record and Annotation Record for a Single Example Image**

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



**Figure 9: Australian Architecture Ontology [59]**