

# A Collaborative Scholarly Annotation System for Dynamic Web Documents - a Literary Case Study

Anna Gerber<sup>1</sup>, Andrew Hyland<sup>1</sup> and Jane Hunter<sup>1</sup>

<sup>1</sup> University of Queensland, St Lucia, Queensland, Australia, (617) 3365 1092  
{agerber, ahyland, jane}@itee.uq.edu.au

**Abstract.** This paper describes ongoing work within the Aus-e-Lit project at the University of Queensland to provide collaborative annotation tools for Australian Literary Scholars. It describes our implementation of an annotation framework to facilitate collaboration and sharing of annotations within research sub-communities. Using the annotation system, scholars can collaboratively select web resources and attach different types of annotations (comments, notes, queries, tags and metadata), which can be harvested to enrich the AustLit collection. We describe how rich semantic descriptions can be added to the constantly changing AustLit collection through a set of interoperable annotation tools based on the Open Annotations Collaboration (OAC) model. RDFa enables scholars to semantically annotate dynamic web pages and contribute typed metadata about the IFLA FRBR entities represented within the AustLit collection. We also describe how the OAC model can be used in combination with OAI-ORE to produce scholarly digital editions, and compare this approach with existing scholarly annotation approaches.

**Keywords:** Annotation, Interoperability, Scholarly Editions, Ontology

## 1 Background

The eResearch Lab at the University of Queensland has been working with the Australian Literature community through the Aus-e-Lit project [1], developing eResearch tools for scholars of Australian literature through the AustLit web portal. AustLit [2] is a collaboration led by the University of Queensland, between twelve Australian universities and the National Library of Australia. AustLit aims to support scholars undertaking research into Australian literary heritage and print cultures by providing authoritative bibliographic information for Australian creative and critical literature works and a selection of articles, poetry and fiction in full text. The research activities within AustLit are focused around particular topics, regions and genres through Research Communities. AustLit currently supports sixteen different Research Communities, including Black Words, Children's Literature and Popular Fiction.

The AustLit data model is based on the IFLA FRBR [3], extended with event modeling. Each AustLit *work* record presents bibliographic information relating to a FRBR *Work*, which is realized through *versions* (FRBR *Expressions*), which are embodied by *publications* (FRBR *Manifestations*). AustLit does not record any details about FRBR *Items*. TEI XML full texts are available for selected works. AustLit also

records biographical information about people and organisations (agents) involved in the creation and dissemination of Australian Literature, as well as information about the events in which they participate including work creation, realization, and manifestation, and their relationships with other agents and works.

Sub-communities use their own specific vocabularies and terminologies associated with their topic of interest. If they wish to contribute their research data and knowledge to AustLit, they need to request that additional attributes are added to the AustLit data model by technical staff. This process presents a delay and barrier to establishing new topics and has also led to increasing complexity of the AustLit data model. This has also affected the user interface for editing AustLit records, making it inaccessible to scholars who do not have adequate training in information systems. As collaborative research teams have become increasingly geographically distributed, scholars have also identified a need for tools to support discussion, sharing of notes and data on projects such as scholarly editions. An analysis of annotation needs within the AustLit community has indicated there are at least five different use cases for annotation tools:

1. Fine-grained semantic tagging of textual documents (both manual and automated);
2. Subjective attachments of free-text notes and interpretations of resources;
3. Input of metadata descriptions for AustLit resources and derived resources;
4. Representation of different versions of scholarly editions as annotations;
5. Representation of compound objects as annotated links between resources.

Hence, the primary motivation for the work described in this paper was to provide collaborative annotation and scholarly editing tools integrated within the AustLit web portal, to enable scholars to collaboratively select and annotate digital resources, and to share their annotations with the research community and enrich the AustLit collection. The remainder of the paper is structured as follows: Section 2 describes Related Work; Section 3 outlines the objectives of the work described here; Section 4 describes the architecture, implementation and user interface; Section 5 provides a discussion of the results; Section 6 outlines future work plans and Section 7 provides a brief conclusion of the outcomes.

## **2 Related Work**

Web-based collaborative annotation systems are designed to enable online communities of users to attach comments or notes to web resources and to tag them with keywords. An overview of existing web annotation systems is provided in [4]. Existing web annotation systems support the first three use cases described in Section 1, but do not support the use cases for scholarly editions or compound objects. Boot [5] provides a discussion of existing approaches to annotation for scholarly digital editions. Many scholarly edition projects employ bespoke annotation systems, using proprietary or non-standard annotation formats. These approaches typically focus on annotating source documents that are encoded using TEI XML [6], and assume that the source documents are static: when changes are made a new document will be created to represent the new version. TEI allows both inline and stand-off annotations. Inline annotations do not allow annotations to overlap, and also require that the

annotator have write access to the source document, making it difficult to maintain the integrity of the contents, and almost impossible to implement in a collaborative web-based environment. Stand-off annotation allows multiple layers of potentially overlapping annotation hierarchies to be stored; however this approach has limitations for some types of scholarly annotations [7], and does not provide a complete solution for annotating web resources generally, including non-TEI documents or images.

The Open Annotations Collaboration (OAC) [8] has been developing a data model and framework to enable the sharing and interoperability of scholarly annotations across annotation clients, collections, media types, applications and architectures. The OAC model, outlined in Figure 1, draws on initiatives such as the W3C's Annotea [9] and the W3C Media Fragments Working Group. Representing annotations using RDF and OWL enables the annotated resources to become accessible to the larger Semantic Web, including inferencing and reasoning engines.

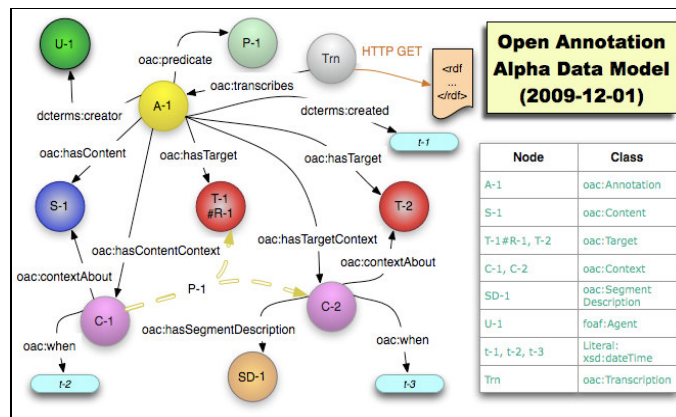


Fig 1. Open Annotations Collaboration model

### 3 Objectives

The first objective of this work was to meet the annotation requirements of the literary scholars using AustLit. Hence, the Aus-e-Lit annotation tools need to support the following activities:

- Scholarly (free text) annotation of textual documents, web pages and images;
- Fine-grained semantic tagging of AustLit documents and web pages (automatic or manual) with user-specified controlled terms (AustLit data entities) to enable search and inferencing;
- Annotation of changes between documents, for example for recording the transmission history of a particular text (for tracking and visualization of scholarly editions).

AustLit scholars also work with documents and images sourced from a variety of institutional repositories, both within national and international collections and on the Web. Hence a further objective was to allow the scholars to use the same annotation system to annotate resources regardless of location. Consequently, we needed to ensure that annotation content was available an open, interoperable format that could

be easily extracted and exported to other formats for re-use. Hence, additional objectives were:

- To evaluate the Open Annotations Collaboration (OAC) model;
- To compare RDF-based approaches with existing (TEI-XML) approaches.

For literary scholarship (and to enable further semantic reasoning), it is important to be able to identify exactly which work, expression, manifestation, agent, and the section of a literary text displayed on a Web page, is under scrutiny. Most Web annotation systems allow a *context* (textual segment or image region) to be specified to identify the section of interest. In practice, the annotated documents are usually HTML Web pages, with XPointer contexts. However, AustLit is a dynamic system - Web pages are dynamically generated from collections that are frequently being updated with additional or corrected information. Using XPointers to represent contexts for a dynamic page is a fragile, unstable approach, as minor changes to the underlying data or stylesheets that render the page can cause the *context* to become invalid. In order to provide a robust Web-annotation system for dynamic pages, we need to be able to distinguish between content and presentation [10] and refer to the content objects rather than the presentation markup in annotation contexts. While it is possible to adopt conventions for HTML IDs to work around this issue, the relationships between identified sections of a dynamic web page and the data entities from which it was generated are not explicit, or able to be generalized across systems. RDFa [11] is a W3C standard for embedding RDF data in (X)HTML. Our hypothesis is that RDFa is ideal for encoding semantic entities within dynamic web pages. Hence our final objective is to evaluate RDFa for semantic tagging of dynamic documents.

#### 4 Architecture, Implementation and User Interface (UI)

Figure 2 shows a high-level view of the Aus-e-Lit annotation system architecture. Researchers annotate pages on the AustLit web portal, or Web resources from other sites using the Aus-e-Lit annotation client, which is implemented as a Firefox browser extension. The annotations are stored as RDF on a separate Annotation server and can be browsed and searched directly via the annotation client. Metadata attached to the annotations can be selectively harvested into the AustLit database via OAI-PMH.

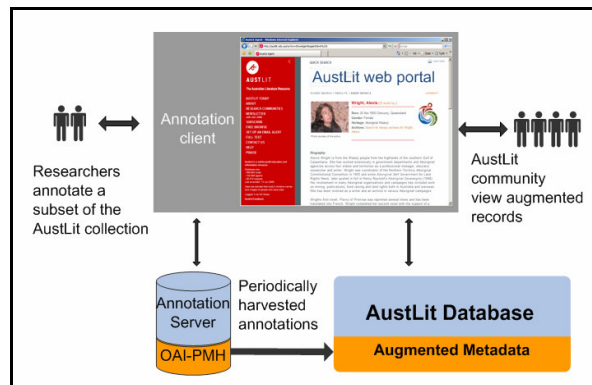


Fig. 2. Architecture diagram

## 4.1 Adding RDFa to the AustLit Web Portal

The AustLit web portal pages have been enhanced with RDFa. The annotation client (described in Section 4.3) can identify, extract and update metadata/tags that are embedded in the AustLit resources. AustLit record pages are rendered from the AustLit database by custom Java servlets using XSLT stylesheets which we extended to insert RDFa into the HTML. The RDFa uses classes and properties from the AustLit ontology that we developed for use with LORE [12] to represent the data entities from which the pages were generated. Figure 3 shows an AustLit work record page alongside a subset of the RDFa that encodes information about the FRBR entities represented on the page.

**AUSTLIT**  
The Australian Literature Resource

**The Drover's Wife** ← **FRBR Work**

SHORT STORY  
Author: Lawson, Henry (birth name: Larsen, Henry (Lawson)) (a.k.a. H. L.)

General subjects: Snakes & serpents, Country life, Women, Dogs, Courage & bravery, Isolation (Emotional & social), Drivers, Families, Fear, Bush

Influence on: [Untitled] - Crisp, Louise

Related To: The Drover's Wife SHORT STORY - Ball, Murray (1979), The Drover's Wife SHORT STORY: SATIRE - HUSACOR - Moorhouse, Frank (1960), The Drover's Wife SHORT STORY: SATIRE - Jeffers, Barbara (1960), The Drover's De Facto SHORT STORY - Gambling, Anne (1980), The Drover's Wife's Dog SHORT STORY - SCHFI - Broderick, Damien (1993), The Bush Undertaker and the Drover's Wife (2006) - Lawson, Henry; Mornit, Stewart (2006)

This work has appeared in at least 4 different versions: ← **FRBR Expression**

- Publications of this version include the following 32: ← **FRBR Manifestation**
  - The Bulletin vol.12 no.649 23 July 1892 PERIODICAL ISSUE (pp.21-22) Find the full text at University of Queensland Library
  - Short Stories in Prose and Verse SELECTED WORK: POETRY PROSE SHORT STORY Author: Lawson, Henry (birth name: Larsen, Henry (Lawson)) (a.k.a. H. L.) Sydney, New South Wales : Louisa Lawson, [1894] (no. 76-38)

```

<div typeof="austlit:Work"
id="CMCg" about="#CMCg"
property="austlit:topicID"
content="CMCg">

<span rel="austlit:hasTitle">
<span typeof="austlit:Title"
id="SPJi" about="#SPJi"
property="austlit:title">
The Drover's Wife
</span></span>

<span rel="austlit:form">
<span typeof="austlit:Form"
about="http://austlit.../ns#i$"
property="austlit:topicName">
short story
</span></span>

...
<span rev="austlit:usedInput">
<span typeof="austlit:Realisation"
id="Z$" Fv" about="#Z$" Fv">

<span rel="austlit:producedOutput">
<span typeof="austlit:Expression"
property="austlit:topicID"
content="Eb7N">

```

Fig. 3. RDFa representing FRBR entities and properties for *The Drover's Wife*

## 4.2 Annotation model

In order to develop a set of annotation tools that will satisfy the five use cases outlined in Section 1, we need a common model that is interoperable across all of them; hence we have adopted the OAC model. Annotations are stored as RDF using Danno [13]; an annotation server developed at the UQ eResearch Lab. Danno uses the Annotea schema to ensure compatibility with existing Annotea clients and servers, while also allowing us to store additional attributes for use by OAC-aware annotation clients. Aus-e-Lit annotations include the following attributes from the OAC model:

- *oac:hasContent* (the URI to the resource containing the content of the annotation – in our case this is an HTML document);
- *oac:hasTarget* (the URI of the web document being annotated);
- *oac:hasPredicate* (indicates the kind of annotation, such as question, comment, explanation, change etc);

Additional information can be recorded using attributes from other schemas such as Dublin Core (*language, format, title, subject, date created or modified* etc). In addition to implementing basic annotation types representing questions, comments

and so on, we extended the model to define a *ScholarlyAnnotation* class, with subclasses *VariationAnnotation* and *SemanticAnnotation* to support the types of scholarly annotations requested by AustLit scholars. Our extensions are as follows:

**ScholarlyAnnotation:** Extends annotations to include *tags*, *importance*, *alternate body* and *references*, based on scholarly annotation requirements outlined in [14].

**SemanticAnnotation:** A subclass of *ScholarlyAnnotation*, which allows scholars to attach metadata conforming to the AustLit ontology to AustLit entities. It extends the model with a semantic context which indicates the entity or property from the web page's RDFa to which the metadata applies.

**VariationAnnotation:** Records metadata about two variants of a text, for example to annotate changes between them. It subclasses *ScholarlyAnnotation* and allows two targets, *variantTarget* and *originalTarget* which are subProperties of *oac:hasTarget*, to identify the original and variant texts. It also adds attributes for recording the *date*, *agent* (person) and *place* where the variation on a text occurred. Table 1 shows the attributes for a *VariationAnnotation*. The namespace *v* represents the Aus-e-Lit *VariationAnnotation* schema.

**Table 1.** Example *VariationAnnotation* record

<b>dcterms:created</b>	2009-09-14T15:57:42.375-07:00
<b>dcterms:modified</b>	2009-09-14T16:11:42.512-07:00
<b>oac:hasPredicate</b>	http://austlit.edu.au/ontologies/2009/03/lit-annotation-ns#VariationAnnotation
<b>dc:title</b>	Editorial Intervention
<b>dc:creator</b>	Roger Osborne
<b>oac:hasContent (stored as separate HTML document)</b>	H. M. Martin was commissioned to edit a collection of poetry from Harpur's manuscript books. The extent of his editorial intervention can be seen in the variation between the 1867 and 1883 versions of 'The Creek of the Four Graves'.
<b>v:variation-date</b>	[1882?]
<b>v:variation-agent</b>	H. M. Martin
<b>v:variation-place</b>	Adelaide
<b>v:variantTarget</b>	http://www.austlit.edu.au/.../1883.txt# xpointer(string-range(/html[1]/body[1]/pre[1], "", 47, 300))
<b>v:originalTarget</b>	http://www.austlit.edu.au/.../MSA871867.txt#xpointer(string-range(/html[1]/body[1]/pre[1], "", 73, 475))

### 4.3 Annotation Client

The Aus-e-Lit annotation client is implemented as an open source add-on to the LORE Firefox extension [12]. The User Interface elements are implemented using HTML and AJAX using the cross-platform ExtJS library, so that it will be possible to port the client to alternative browser extension frameworks in the future. The client provides several views that allow scholars to browse, search and edit annotations:

- A tree-based Browse view, which lists all annotations on a given page, along with any replies. The annotations in the tree can be sorted in ascending or descending order by date of creation, creator or annotation title;
- Timeline view, displaying a summary for all annotations and replies by date;
- Search view, allowing search by creator, title or date-range;
- Edit view, which allows scholars to create or modify existing annotations;

- VariationAnnotation view, which shows resources that are being compared or related side-by-side along with additional scholarly metadata.

Figure 5 shows the Aus-e-Lit Annotation client. The main browser window displays the VariationAnnotation view, which relates a manuscript image from the notebook of Patrick White with a transcript of another version of the text. The Browse view is visible in the top-right, and the Edit view is shown at the bottom-right of screen, including fields for editing the additional scholarly annotation attributes that were described in 4.2. For example, the Tags field allows scholars to tag resources using keywords, which can be selected from the AustLit thesaurus, or entered as free text. The user interface automatically suggests matching tags as the user types. The list of fields displayed in the editor changes depending on the type of the annotation.

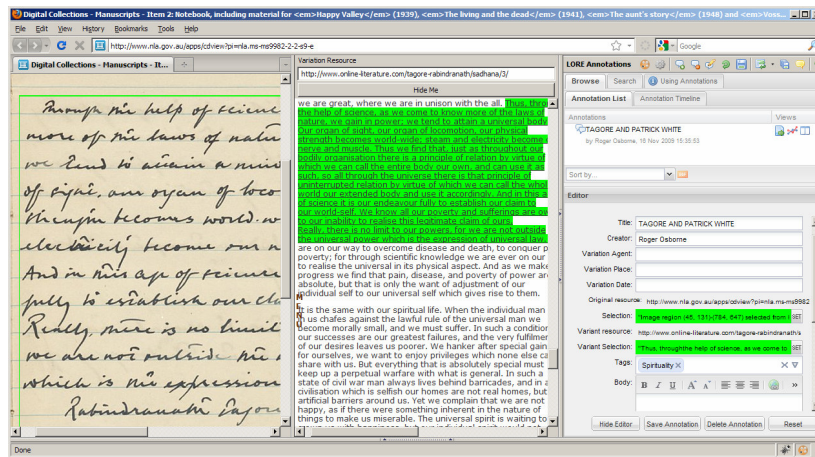


Fig 5. Editing a VariationAnnotation relating a manuscript image and transcript

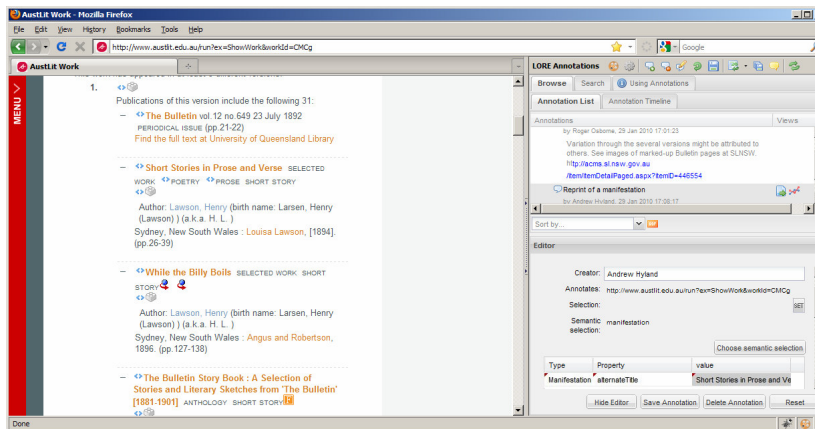


Fig 6. Editing a SemanticAnnotation that attaches metadata to a Manifestation

A key objective was to make semantic annotations accessible to scholars by providing a user interface that hides the complexity of the AustLit data model and does not require a detailed understanding of IFLA FRBR. We designed the UI to allow scholars to attach metadata directly to entities such as works or manifestations via the familiar AustLit Web interface. In Figure 6, an *alternateTitle* property is being contributed for an AustLit manifestation via a SemanticAnnotation. When the *Choose semantic selection* button is pressed, the annotation client parses the embedded RDFa and inserts icons into the Web page; a brick icon next to each entity, and a brackets icon next to each property from the RDFa. The user selects the entity that they wish to annotate by clicking on these icons. A drop down menu in the editor allows users to attach OWL Data Type properties from the AustLit ontology to the annotation. The UI only displays properties that are applicable for the selected entity, and basic validation is performed on the data values entered, to assist scholars to enter data that conforms to the AustLit data model.

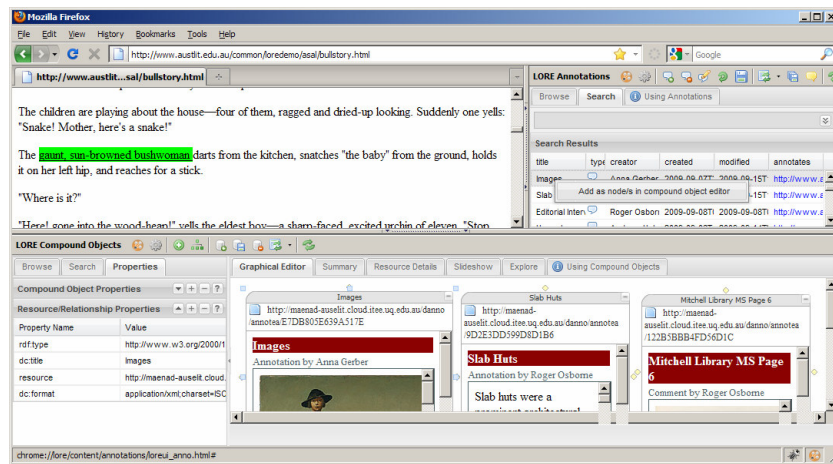


Fig 7. Publishing a scholarly digital edition using LORE

LORE [12] allows metadata and relationships to be specified at the resource level, which are then published as OAI-ORE-compliant RDF. Annotations extend the capabilities of LORE so that parts of resources (such as image regions, or sections of a text) can be discussed, and annotations can be attached to this context. As each annotation is itself a resource with a unique URI, a Scholarly Edition can be built up by a group of collaborators annotating multiple documents representing different versions of a work, and then published as a LORE compound object that encapsulates all of the versions, along with *ScholarlyAnnotations* that provide critical commentary and *VariationAnnotations* that describe each variation in detail. Figure 7 shows how the annotation client integrates with LORE to allow annotations to be selectively added to a compound object from the annotation search or browse panel. In addition to publishing collections of annotations using LORE, scholars can select annotations for export to RDF/XML or to a Microsoft Word document, which lists the metadata, target URI(s) and contents of the selected annotations.

## 5 Discussion

A survey of existing scholarly annotation systems [15] revealed that the majority:

- focus on scholarly formats such as TEI to the exclusion of general Web resources;
- are closed systems that do not allow annotation of outside documents;
- act as silos: annotations are not easily reused or referenced outside of the system.

This presents a problem for collaboration and re-use of annotations. The OAC model allows any online resource to be the target or content of an annotation, regardless of media type or location, and is designed to facilitate interoperability between annotation systems. Our case study has demonstrated that the OAC model can be extended to enable specialized Scholarly Annotations that record scholarly metadata and relate multiple documents. Web pages are often discounted as scholarly resources as they are considered to be “transitory representations of the scholarly objects that need Annotation” [5]. By using RDFa to encode the relationship between presentation markup and the scholarly or data entities represented, our system enables scholarly annotation of dynamic Web pages from digital library and information systems.

Feedback from AustLit scholars has included a request for semantic contexts for *VariationAnnotations*. While our semantic context works well for AustLit record pages, it remains to be seen how well this approach generalizes to Web pages representing full texts rendered from TEI XML. We believe our approach can be generalized to work with other ontologies and types of data, however, the following considerations should be noted: Current JavaScript libraries for extracting and working with RDFa are not very efficient, however we expect that browser support will improve for RDFa over time. Each data entity of interest needs a URI. Blank nodes should be avoided in RDFa, while triples need to be reified in order to be identified within the semantic context. Our system generates a unique hash that identifies the context triple, however a better approach would be to construct Named Graphs. XPointer contexts are also used in combination with the semantic context for presentation markup that exists inside of RDFa spans. This is necessary because the granularity of our RDFa does not allow us to specify contexts down to the character level. This will be a problem when working with TEI documents, as the sections within AustLit’s TEI documents are very large. We may need to enrich the original TEI documents to provide finer section granularity.

## 6 Future Work

Further work to be undertaken on this project includes a detailed performance and user evaluation of the system and the following:

- To improve the integration of the annotation client and the LORE Compound Object editor for publishing scholarly digital editions, including options to export scholarly compound objects that contain annotations to TEI documents;
- To store versioning information for Web documents in order to alert scholars about the changes that may have occurred since the page was annotated;
- To harvest and map metadata contributed through annotation into AustLit, with a facility for filtering and moderating the annotations for inclusion;
- To investigate using OAI-ORE with the OAC model for Variation annotations so that users can compare and record changes between more than two documents.

## 7 Conclusions

In this paper we have described the implementation of a scholarly annotation and publishing framework that enables literary scholars to annotate web resources with tags, comments and notes using the Open Annotations Collaboration (OAC) model. We have extended the OAC model to enable scholars to create *VariationAnnotations* which document changes between texts, and *SemanticAnnotations* for contributing typed metadata. Our annotation client leverages RDFa embedded within dynamic Web pages to identify and to allow annotations to be associated directly with the underlying data entities, rather than with the constantly changing presentation markup. Finally, through integration with LORE, annotations and other digital resources can be aggregated into OAI-ORE compound objects representing scholarly digital editions, and shared with the research community to enrich the AustLit collection.

**Acknowledgements.** Aus-e-Lit is funded by DEST through the National eResearch Architecture Taskforce. We gratefully acknowledge the valuable contributions made to this paper by Roger Osborne, Kerry Kilner and the AustLit research communities.

## References

1. Aus-e-Lit, <http://www.itee.uq.edu.au/~eresearch/projects/aus-e-lit/>
2. AustLit: The Australian Literature Resource, <http://austlit.edu.au>
3. Kilner, K.: The AustLit Gateway and Scholarly Bibliography: A Specialist Implementation of the FRBR. *Cataloging & Classification Quarterly*, vol. 39, no. 3/4. (2004)
4. Hunter, J., Khan, I., Gerber, A.: HarVANA - Harvesting Community Tags to Enrich Collection Metadata, JCDL. PA, USA, June 16 – 20, pp 147 - 156 (2008)
5. Boot, P.: Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship. Pallas Proefschriften, Amsterdam. PhD Thesis. (2009)
6. Text Encoding Initiative (TEI), <http://www.tei-c.org/index.xml>
7. Banski, P., Przepiorkowski, A.: Stand-off TEI Annotation: the Case of the National Corpus of Polish, ACL-IJCNLP LAW III, Singapore, August 6 –7 (2009)
8. Open Annotations Collaboration, <http://www.openannotation.org/>
9. W3C Annotea Project, <http://www.w3.org/2001/Annotea/>
10. Hemminger, B.: NeoNote. Suggestions for a Global Shared Scholarly Annotation System. *D-Lib Magazine*, May/June (2009)
11. W3C RDFa in XHTML, <http://www.w3.org/TR/rdfa-syntax/>
12. Gerber, A., Hunter, J.: LORE: A Compound Object Authoring and Publishing Tool for the Australian Literature Studies Community, In: ICADL 2008. LNCS, vol. 5362, pp. 246--255, Springer, Berlin /Heidelberg (2008).
13. Chernich, R., Crawley, S., Hunter, J.: Universal Collaborative Annotations with Thin Clients - Supporting User Feedback to the Atlas of Living Australia, eResearch Australasia, Sydney, Australia, November 13 – 15 (2009)
14. Furuta, R., Urbina, E.: On the Characteristics of Scholarly Annotations. Conference on Hypertext and Hypermedia, MD, USA, pp 78 – 79 (2002)
15. Hunter, J.: Collaborative Semantic Tagging and Annotation Systems, In: Annual Review of Information Science and Technology, American Society for Information Science & Technology, vol. 43 (2009)