Submitted for consideration for DH2010 conference issue of **Literary and Linguistic Computing**

The Open Annotation Collaboration:

A Data Model to Support Sharing and Interoperability of Scholarly Annotations

Timothy W. Cole

University of Illinois at Urbana-Champaign, Urbana, IL, USA

Jane Hunter

The University of Queensland, Brisbane St Lucia, QLD, Australia

Robert Sanderson and Herbert Van de Sompel

Los Alamos National Laboratory, Los Alamos, NM, USA

Correspondence:

Timothy W. Cole

t-cole3@illinois.edu

216 Altgeld Hall (MC-382)

1409 W. Green St.

Urbana, IL 61801

USA

The Open Annotation Collaboration:

A Data Model to Support Sharing and Interoperability of Scholarly Annotations

Abstract

This paper presents the outcomes to date of the in-progress Open Annotation Collaboration (OAC) Project.[1] The members of the OAC, the University of Illinois, the University of Queensland, Los Alamos National Laboratory Research Library, George Mason University and the University of Maryland, have received funding from the Andrew W. Mellon Foundation to develop 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 Web and resource-centric approach to annotation description is based on the assumption that clients publish annotations on the Web and that the annotation target, annotation body, and the annotation itself are all instantiated as URI-addressable Web resources. By basing the OAC data model on Semantic Web and Linked Data practices, our goal is to provide the optimum approach for the publishing, sharing and interoperability of annotations and annotation applications. In this paper, we describe the guiding principles adopted for this work, the current release of the OAC data model for interoperable annotations, and a sampling of the use case-driven issues and questions on which we are relying to test, refine, and assess the OAC data model across a range of different scholarly scenarios.

1 Introduction and Objectives

Annotating, the act of associating one piece of information with one (or more) other piece(s) of information, is both a core and a pervasive practice for humanities scholarship and for scholarship generally. It is used to organize, create and share knowledge. Individual scholars use it when reading, as an aid to memory, to add commentary, and to classify documents. It can facilitate shared editing, scholarly collaboration, and pedagogy. Although there exists a plethora of annotation clients for humanities scholars to use (Hunter 2009), many of these tools are designed for specific collection types, user requirements, disciplinary application or individual, desktop use. Scholars are confronted with having to learn different annotation clients for different content repositories, have no easy way to integrate annotations made on different systems or created by colleagues using other tools, and are often limited to simplistic and constrained models of annotations suitable for use only in specific contexts. For example, many existing tools only support annotation models which restrict annotation content to a brief unformatted text string. Equally problematic are the many tools that fail to treat annotations as first-class, independent information objects, instead conflating the storage of the annotation and the target resource being annotated. This approach can frustrate subsequent efforts to directly reference or annotate an annotation in its own right. To help address these and related issues, the objectives of the OAC are:

- To facilitate the emergence of a Web and resource-centric interoperable annotation environment that allows leveraging annotations across the boundaries of annotation clients, annotation servers, and content collections.
- To demonstrate, through prototype implementations, an interoperable annotation

environment in a variety of settings characterized by a range of annotation client/server configurations, content collections, and scholarly use cases.

- To seed widespread adoption of OAC standards in scholarly contexts by deploying robust, production-quality applications conformant with this interoperable annotation environment.

2 Related Work and the OAC Guiding Principles

Despite the vast body of work regarding annotation practice, annotation models, and annotation systems, relatively little attention has been paid to interoperable annotation environments. The few efforts in this realm to date comprise:

- RDF-based *Annotea* developed by Kahan and Koivunen (Kahan et al., 2001);
- Agosti's *Formal Model of Annotations of Digital Content* (Agosti and Ferro, 2007);
- Boot's *SANE: Scholarly Annotation Exchange*, based on a model of third-party annotations presented at DH2006 (Boot, 2006)
- *OATS: The Open Annotation and Tagging System* (Bateman et al., 2006).

An analysis of these existing models reveals that on the whole, they have not been designed as Web-centric and resource-centric and/or that they have modeling shortcomings that prevent any existing resource from being the content or target of an annotation and preclude an annotation from being given independent status as a resource itself. To help ensure an advance on prior art, the OAC has articulated a set of guiding principles[2] which inform our work, including:

- The OAC effort will focus on annotation interoperability; we will test the OAC data model using both existing and special-purpose annotation clients, but will not prescribe client user interface design or internal client architecture.
- Contrary to views that restrict the body (content) of an annotation to simple text, we hold

that annotation bodies (as well as targets) may be of any media type.

- Annotations, as well as the bodies and targets of annotations, should all be (HTTP) URI-addressable resources.
- Many annotations involve parts of resources (image regions, slides of a video, text fragments); therefore, support must be provided for addressing segments of resources.
- An annotation may comprise one or more body resources and one or more target resources.
- URI-addressable resources are ephemeral: the representations obtained by dereferencing their URIs may change over time; therefore, support must be provided for persistent annotations (Sanderson and Van de Sompel, 2010).
- While providing a set of top-level classes/entities and properties/relationships that will maximize interoperability across annotation clients, servers, collections and applications, the ontology of our shared annotation data model must be extensible.

3 The OAC Data Model

A preliminary version of a data model to support annotation sharing and interoperability has been released by the OAC.[3] This data model provides a method of describing annotations such that they can easily be shared between platforms, with sufficient richness of expression to satisfy scholars' needs while remaining simple enough to also allow for common use cases such as attaching a piece of text to a single web resource. Consistent with practices encouraged by the Linked Data Initiative[4], annotations are modeled as a set of connected (HTTP) URI-addressable resources, including one or more *annotation body* resources, i.e. the annotation content or source, and one or more *annotation target* resources. Figure 1 depicts the baseline (simplest) instantiation of the OAC data model represented as a graph of connected resources and properties.

While an essential aspect of any annotation is the (implicit or explicit) expression of an

annotates relationship between the body and target, the approach of treating the annotation, the body, and the target as distinct, URI-addressable resources simplifies and decouples implementation from the perspective of the repository. The OAC data model allows for the body and target to be of any media type. Moreover the model allows the annotation, the body, and the target each to be stored separate one from another, and for the annotation, the body, and the target to all have different authorship. This latter feature supports ontological-based annotation, i.e. the annotation of resources using concepts drawn from pre-existing discipline-specific ontology (e.g. Shah et al., 2009).

To facilitate interoperability of annotation applications and to leverage existing Semantic Web standards and practices, the OAC data model allows for the expression of additional properties and relationships involving the annotation, body, and/or target. In particular, the annotation itself will have additional properties and relationships as illustrated in Fig. 2.

While the OAC data model avoids conflating the annotation body with the annotation, often the instantiation of the annotation and the annotation body occur simultaneously, e.g. the annotation comes into being at the same time as the textual comment which is the body of the annotation. In such cases it may be convenient to instantiate the annotation body inline within the annotation serialization. The OAC data model leverages the W3C's *Representing Content in RDF* Working Draft[5] to allow this in a manner consistent with Semantic Web best practice. For example, a text annotation body can be embedded within an annotation serialization as illustrated by the graph depicted in Fig. 3. The *Representing Content in RDF* Working Draft also supports XML (including XHTML) and base64-encoded annotation bodies. To create the URI of an inline annotation body, a *urn:uuid* may be used.

Scholarly annotations are often more complex than the annotation descriptions illustrated

in Figures 1-3. Annotations that compare and contrast resources involve multiple annotation targets. Some scholarly annotations may involve a body or a target that is a specific representation of a resource or a segment or fragment of a particular resource. Scholarly annotations may involve annotation body or target resources that are ephemeral, subject to change, or have other time dependencies. The OAC data model provides features by which these more complex annotations can be described. The use of URIs including fragment identifiers is encouraged. In order to allow for use cases which cannot be described using fragment URIs alone, the OAC data model defines two additional entities, the *ConstrainedTarget* and the *ConstrainedBody*, and 2 special predicates, *constrains* and *constrainedBy*. As illustrated in Fig. 4, these entities and relationships can be used to describe annotations that target (or have as body) a specific segment, representation, or version of a resource. (The ConstrainedTarget in Fig. 4 is the node identified as 'uu1.')

4 Use Cases and Questions to be Investigated

To inform and support the development of the OAC data model, a range of scholarly annotation use cases were examined. These initial use cases were developed from a review of the literature, augmented by direct discussions with scholars in multiple disciplines. Annotations involving both digital and non-digital resources were examined.

For example, the process of developing a scholarly edition involves analyzing and annotating multiple editions of a specific work. These tasks may be carried out by one individual, or as is often the case today, by a group of scholars. Increasingly, as more and more retrospective digitization projects come online, much of this work can be done using digital surrogates of published editions and variants. Figure 5 illustrates annotations pertaining to differences between versions of literary works. The left-hand portion of Fig. 5 shows annotations made against digital surrogates of The Creek of Four Graves, a poem by Charles Harpur; the right-hand portion of Fig.

5 shows hand-written annotations of a print copy of *An Old Mate of Your Father's*, a short story by Henry Lawson serialized in the newspaper, *The Bulletin*, and later included in a short story collection published in multiple editions.

Figure 6 shows how such a complex annotation, in this case an annotation involving multiple resources as components of a compound or aggregated target, can be modeled using the OAC data model in conjunction with the *Open Archives Initiative Object Reuse and Exchange* (OAI-ORE) specification.[6]

Our initial examination of scholarly annotation use cases has raised other data modeling issues and questions beyond those mentioned above. Many of these questions remain to be resolved. Among them:

- Through HTTP content negotiation it is sometimes possible to dereference a given URI so as to retrieve a preferred representation of a resource, e.g. to retrieve an image in one format rather than another, or a text in French rather than English. The OAC data model provides a means to reference a preferred representation of a given resource as annotation body or target. How often is this functionality required in practice?
- As illustrated above, the OAC data model accommodates annotations involving multiple targets as well as annotations targeting OAI-ORE Aggregations. What factors should implementers consider when deciding which approach to use?
- Work to date suggests that issues of anchor vs. citation may be important in some scholarly domains. A literary scholar may annotate a passage in a novel while viewing a specific digital instance of that novel. While the annotation target can be anchored in the specific digital instance, often the intent is to annotate the passage in multiple digital instances of the novel, possibly spanning editions of the work. Can such intent be expressed using the

OAC data model?

- An annotation may both target one resource and reference another resource. How is this use case distinguished from an annotation involving multiple annotation targets?
- Biodiversity annotation use cases examined suggest that there are times when a user may wish to annotate a resource in context, e.g. a scholar may want to say something about an image, but only about the image as it is embedded in a specific Web page. How important a use case is this in other domains, and if important, is it well enough handled by the OAC data model.

5 Discussions and Conclusions

The proposed OAC Data Model will enable the sharing and discovery of annotations beyond the boundaries of individual solutions and content collections, and hence will allow for the emergence of value-added, cross-environment annotation services. It also will facilitate the implementation of advanced end-user annotation services that are capable of operating across a broad range of both scholarly and general collections. Furthermore, it will enable customization of annotation services for specific scholarly communities, without reducing interoperability, and will enable more robust machine-to-machine interactions and automated analysis, aggregation and reasoning over distributed annotations and annotated resources. By grounding this work in a thorough understanding of Web-centric interoperability and embedded models implemented by existing digital annotation tools and services, we create an interoperable annotation environment that will allow scholars and tool-builders to leverage prior tool development work and traditional models of scholarly annotation, while simultaneously enabling the evolution of these models and tools to make the most of the potential offered by the Web environment.

Funding and Acknowledgments

This work was supported by the Andrew W. Mellon Foundation. The authors would also like to acknowledge the valuable contributions to this work made by: Neil Fraistat, Doug Reside, Daniel Cohen, John Burns, Anna Gerber, Tom Habing, Clare Llewellyn, Carole Palmer, Allen Renear, Bernhard Haslhofer, Ray Larsen, Cliff Lynch and Michael Nelson.

References

Agosti, M. and Ferro, N. (2007). A Formal Model of Annotations of Digital Content. *ACM Transactions on Information Systems*. 26 (1): online. Available: http://dx.doi.org/10.1145/1292591.1292594 [accessed 8 October 2010].

Bateman, S., Farzan, R., Brusilovsky, P., McCalla, G. (2006). *OATS: The Open Annotation and Tagging System, Proceedings of the Third Annual International Scientific Conference of the Learning Object Repository Research Network*, Montreal, November. Available: http://www.cs.usask.ca/~ssb609/files/oats-lornet.pdf [accessed 8 October 2010].

Boot, P. (2006). *Third-Party Annotations in the Digital Edition Using EDITOR, Proceedings of Digital Humanities 2006 [July]*. Paris: Université Paris - Sorbonne, pp. 34-35.

Hunter J. (2009). Collaborative Semantic Tagging and Annotation Systems. *Annual Review of Information Science and Technology (ARIST)* 43: 187 - 239.

Kahan, J., Koivunen, M., Prud'Hommeaux, E., and Swick, R. (2001). *Annotea: An Open RDF Infrastructure for Shared Web Annotations, Proceedings of the 10th International conference on the World Wide Web*, Hong Kong, May. Available: http://www10.org/cdrom/papers/488/index.html [accessed 8 October 2010].

Sanderson, R. and Van de Sompel, H. (2010). *Making web annotations persistent over time, Proceedings of the 10th Annual Joint Conference on Digital Libraries*, Gold Coast,

Queensland, Australia, June. New York: ACM. Available: http://dx.doi.org/10.1145/1816123.1816125 [accessed 11 October 2010].

Shah, N. H., Jonquet, C., Chiang, A. P., Butte, A. J., Chen, R., and Musen, M. A. (2009). Ontology-Driven Indexing of Public Datasets for Translational Bioinformatics. *BMC Bioinformatics* 10 (Supplement 2):S1. Available: http://dx.doi.org/10.1186/1471-2105-10-S2-S1 [accessed 11 October 2010].

Notes

1 http://www.openannotation.org/

2 http://www.openannotation.org/documents/OAC_GuidingPrinciples_20091106.pdf

3 http://annotation.lanl.gov/

4 http://esw.w3.org/LinkedData

5 http://www.w3.org/TR/Content-in-RDF10/

6 http://www.openarchives.org/ore/

Figures

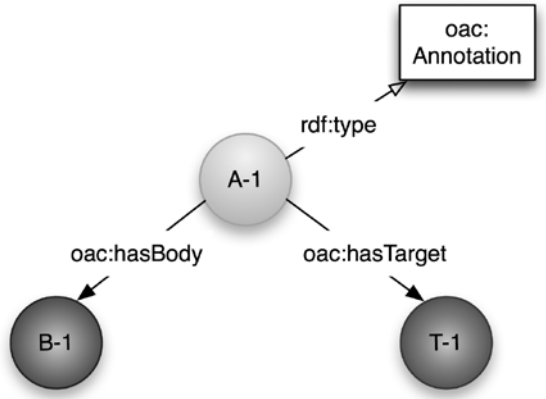

Fig. 1: The baseline OAC data model (Alpha-3 version)

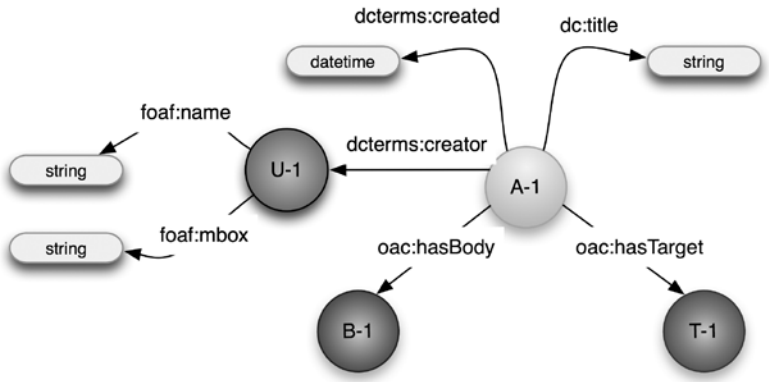

Fig. 2: Associating additional properties and relationships with an annotation

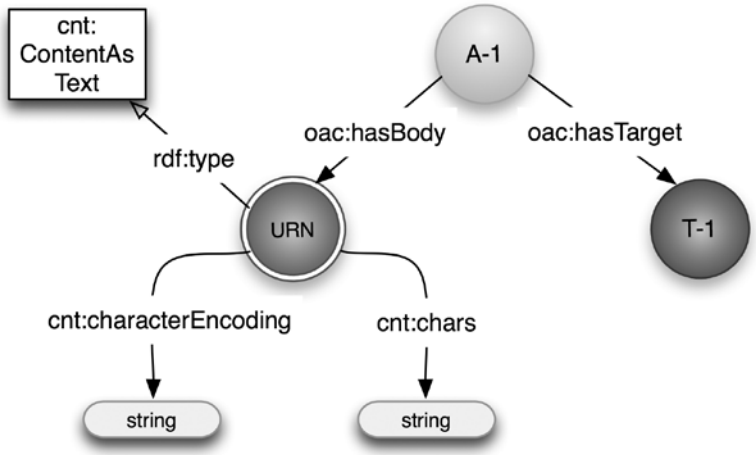

Fig. 3: Annotation with an inline body

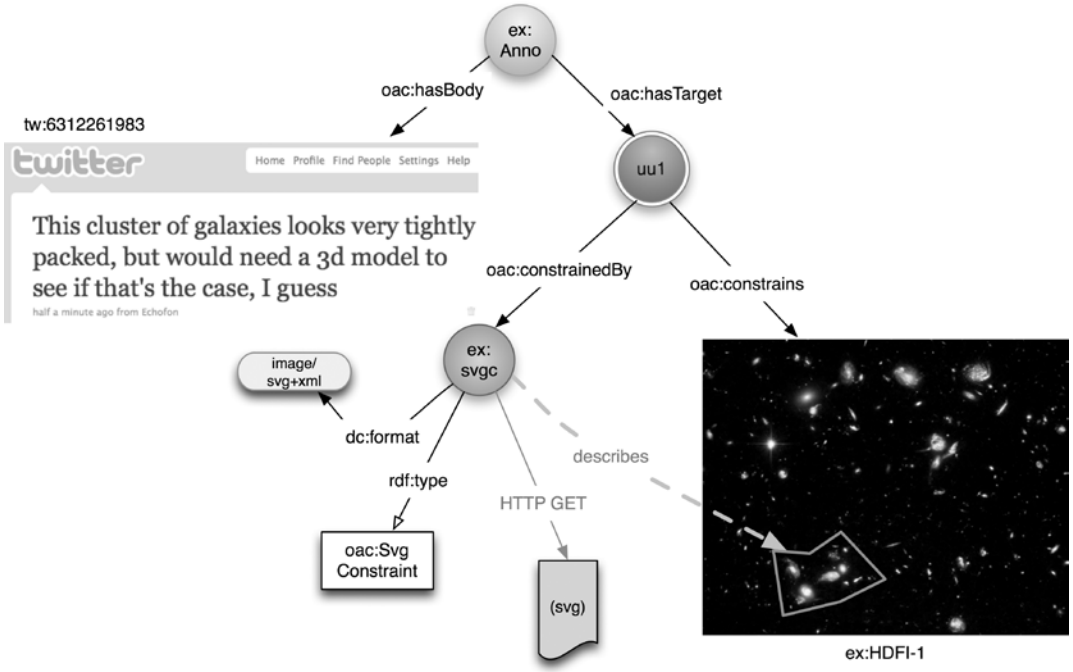

Fig. 4: Using a ConstrainedTarget to reference a segment of a resource as annotation target

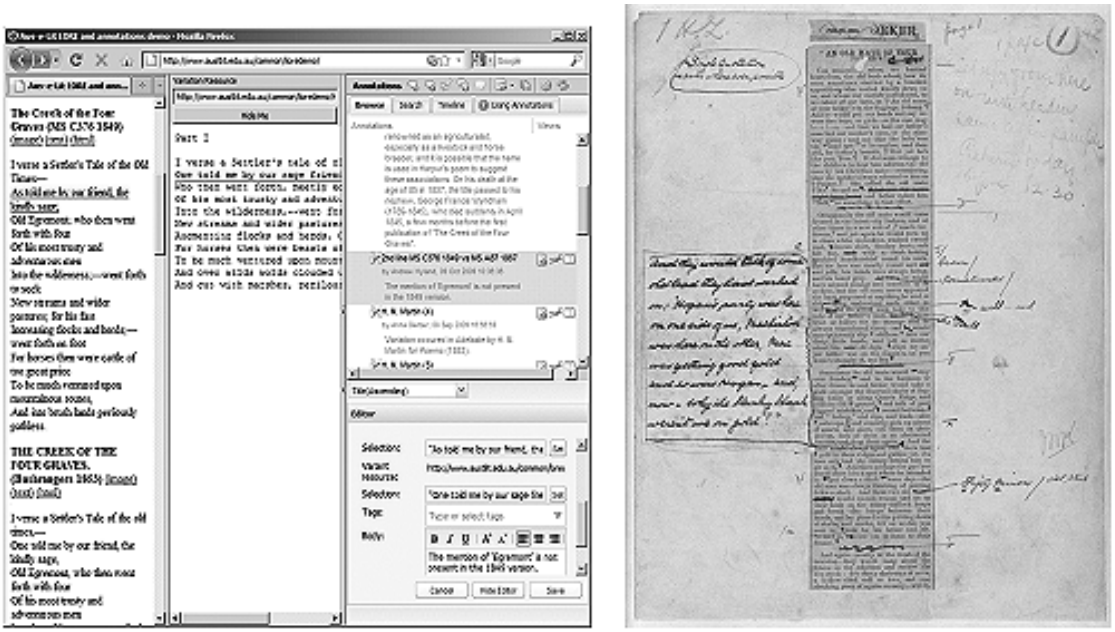

Fig. 5: Annotations detailing differences between published versions of literary works

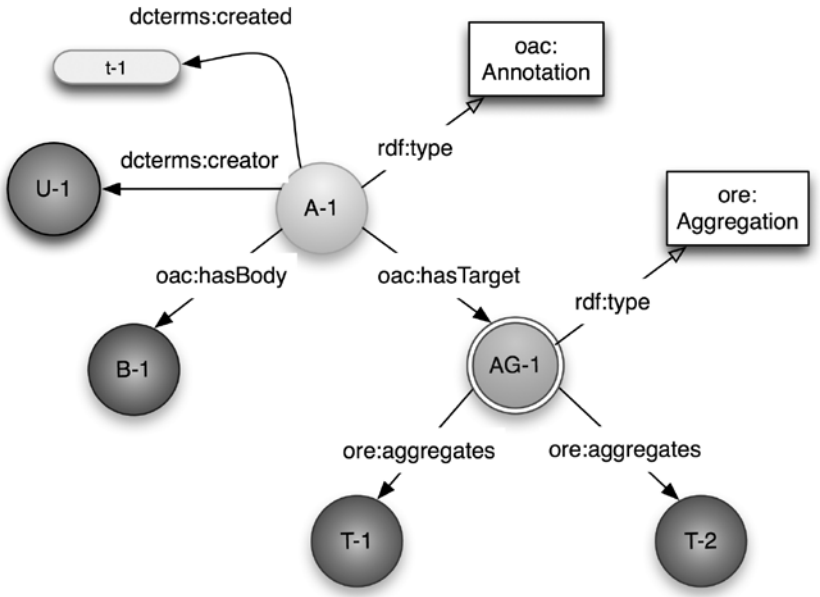

Fig. 6: Graph of an annotation having an OAI-ORE Aggregation as its target