Open Annotation Collaboration Project Review Meeting


Table of Contents

  Summary
  In Attendance
  General Context
  Data Model Q & A Session
  Project Presentations:
  Annotation of Digitized Medieval Manuscripts
  CATCHPlus Open Document Annotation (CODA)
  An early prototype for surfacing many annotations on arbitrary documents
  Annotation Supporting Collaborative Development of Scholarly Editions
  Developing an Annotation Framework for Fedora using OAC
  Automated Annotation of Biomedical Literature
  Annotation of Digital Emblematica
  MapHub: A Demonstrator for Historic Map Annotations
  Annotation of Subscription Streaming Video Content
  Annotation Middleware for Scholarly Publications and Resources
  Conclusions and Next-Steps

Summary

On 26-27 July 2012, members of the Open Annotation Collaboration met in Chicago to present the annotation demonstration experiments that they had been developing to one another and a select group of invited experts. The overall goal of this meeting was to discuss the successes and failures of the OAC data model in the context of the nine experiments. Special guest Dan Whaley also presented on his non-profit company’s use of the emerging Open Annotation specifications. Several common themes surfaced during the discussions and members of the Collaboration agreed to start issue briefs in the Open Annotation Community Group’s webspace to prepare the larger Community for the September Community Group Meeting. These issues included but were not limited to:

- Multiple Bodies,
- Target resource context,
- Scaling of selectors,
- Versioning, etc.
In Attendance
Andrew Ashton (Brown University), Hennie Brugman (Meertens Institute), Paolo Ciccarese (MIND Center for Interdisciplinary Informatics), Tim Cole (University of Illinois at Urbana-Champaign), Alberto Ortiz Flores (New York University Libraries), Anna Gerber (The University of Queensland), Bernhard Hashofer (Cornell University), Nancy Ide (Vassar College), Jacob Jett (University of Illinois at Urbana-Champaign), Randall Leeds (Hypothes.is), Kevin Livingston (University of Colorado Anschutz Medical Campus), Karen Miller (Northwestern University Library), David Millman (New York University Libraries), Mark Notess (Indiana University), Bill Parod (Northwestern University Library), Robert Sanderson (Los Alamos National Laboratory), Jordan Vannoy (University of Illinois at Urbana-Champaign), Dan Whaley (Hypothes.is), Harlan Wallach (Northwestern University)

General Context
The Open Annotation Collaboration Phase II project began in January 2011 with the overarching goal of developing a stable, production level data model and ontology with which digital annotation tool developers could make interoperable tools. The initial Alpha 3 OAC data model was an RDF-based data model produced during OAC Phase I. The Phase II project was charged with refining the data model through experimentation. The project began with four participating annotation demonstration experiment projects:

- **Annotation Supporting Collaborative Development of Scholarly Editions.** Partner: AustLit
- **Annotation of Digitized Medieval Manuscripts.** Partner: DMSTech (lead by Stanford University Library)
- **Annotation of Subscription Streaming Video Content.** Partner: Alexander Street Press / MITH
- **Annotation of Digital Emblematica.** Partner: Emblematica Online (Illinois / Herzog August Bibliothek, Wolfenbüttel)

In March 2011 OAC held a *Using OAC Workshop* in Chicago. The purpose of the workshop was twofold. First, the March Chicago meeting allowed OAC investigators to engage with the larger community of annotation tool developers and stakeholders to promote awareness of the data model and solicit feedback on its shortcomings. Second, the March Chicago meeting provided OAC investigators a venue to announce their Request For Proposals (RFP) to recruit four additional projects to work with the data model.

Response to the RFP was extremely strong and rather the decisions to select the strongest candidates from the respondents were difficult. Ultimately, OAC investigators, in consultation with the Mellon Foundation, agreed to expand the number of projects recruited through the RFP from four to five. The successful respondents were:

- **Annotation Middleware for Scholarly Publications and Resources.** Partner: NYU Libraries
- **Automated Annotation of Biomedical Literature.** Partner: University of Colorado Denver

---

3. [http://www.openannotation.org/documents/openAnnotationRFP.pdf](http://www.openannotation.org/documents/openAnnotationRFP.pdf)
The OAC Beta specification[^4] was published in August 2011 and the nine annotation demonstration experiments began building prototype implementations with the data model.

Concurrent to the ongoing experiments, OAC investigators and investigators at the Annotation Ontology project began holding discussions to align their mutual efforts to design a linked-data compliant, RDF-based interoperable annotation data model. Key members of each project met in Albuquerque, New Mexico, in September 2011. Discussions at that meeting confirmed a strong desire by both groups align their data models.

On the advice of Ivan Herman at the W3C the projects decided to launch a joint W3C Community Group to help facilitate the production of a unified data model. OAC project leaders were able to secure generous supplemental funding to support two face-to-face meetings of the resulting Open Annotation Community Group[^5]. The first of these meetings was held in Boston in March 2012. Initial drafts of a joint Core Specification[^6] and an Extensions Specification[^7] were produced in May 2012, incorporating the group’s decisions at the Boston Meeting.

Following the Boston Meeting, several of the projects experimenting with the OAC Beta specification changed their experiments to accommodate the emerging OA Core & Extensions Specifications. The results of OAC experiments with both the Beta and Core specifications were presented by the nine participating OAC partner projects at the July Project Review Meeting and are summarized below.

In addition to the nine participating project presentations, Dan Whaley, founder of Hypothes.is, demonstrated the Hypothes.is implementation of the OA Core specification. Conceived as a web resource reputation service, Hypothes.is is building digital annotation tools that will allow web-based documents to be peer-reviewed. They are interested in using the Open Annotation specification to facilitate these activities.

### Data Model Q & A Session

The meeting began with a general overview[^8] of the OAC project (provided by Cole) and an update on the status of various project deliverables (provided by Jett). Sanderson then provided an overview of the Open Annotation Core and Extension Specifications.[^9] Among some of the new features in the Extensions document were offset text selectors (Figure 1) and quotation text selectors (Figure 2). Afterwards, a question and answer session was held.

[^5]: [http://www.w3.org/community/openannotation/](http://www.w3.org/community/openannotation/)
I love this passage with the Cheshire Cat!

Figure 1: Offset Text Selectors

I love this passage with the Cheshire Cat!

Figure 2: Quotation Text Selector
During the Q & A session several issues were brought forward by meeting attendees. These issues included:

- **Annotation Typing** – Specifically how can the community avoid overloading classes and class proliferation; use of motivations and expectations in place of, or in addition to, annotation sub-classes was suggested, as was using semantic tags to refine the sub-classes through genre faceting.
- More details and examples regarding the use of the `oa:hasState` was requested. How to make implementations using HTTP headers to aid in content negotiation?
- How the multiple targets case should be interpreted was broached. Are these alternate targets, separate targets that the annotation body applies equally to, or composite targets of some kind? The ambiguity of this case makes reasoning difficult.
- The question of whether or not the object of `oa:hasGenerator` is an application or a software library was raised. The answer given was that it could be either and that this should be left to the implementing community to decide.
- Use of string literals in the Body was raised and it was noted that they are disallowed by the data model because the RDF community does not favor their use (despite explicitly allowing their use in RDF).
- Layering of annotations was brought up as a key use case in the text mining and linguistics fields.
- Finally, the `oa:hasStyle` property was again brought up. In addition to questions about how to apply “styling” beyond CSS (e.g. XSLT), its impact on queries involving Specified Targets was brought up.

**Project Presentations:**

This section will refer to powerpoint slidesets (attached), will highlight specific points made during the presentations, and will detail questions and ideas raised by the presenters and audience.

1. **Annotation of Digitized Medieval Manuscripts\(^{10}\)**

   Sanderson presented the annotation work taking place at Stanford and Los Alamos National Laboratory (LANL). They have implemented a canvas system that allows scholars to view images of digitized manuscripts alongside their transcriptions. Shared Canvas makes use of the OA Data Model in two ways. First, annotations are used as vectors to link the images and text to the specific canvas areas. Second, there are a set of annotation tools that permit scholars using Shared Canvas to create discourse annotations.

   Questions were raised about using annotations to link images and text to the canvas. Isn’t this an example of multiple bodies (which are disallowed by the data model)? The answer to this raised question is that they are in essence multiple bodies but Shared Canvas is modeling them as a single Structured Body to get around this.

---

\(^{10}\) [http://www.openannotation.org/documents/oac2-wkshp2-sharedcanvas.pdf](http://www.openannotation.org/documents/oac2-wkshp2-sharedcanvas.pdf)
2. **CATCHPlus Open Document Annotation – CODA**¹¹

Brugman presented the CODA project, which operates in an environment of federated science and humanities resources. Their need was for an annotation standard that was broad enough to accommodate the needs of multiple scholarship domains and produce annotations interoperable enough that a single tool set can be used to manage them.

One of the interesting use cases that the CODA implementers discovered during the implementation process was the need to identify named entities, and the need for the annotation to link parts of the named entity across multiple lines in a specific sequence. Other issues the CODA implementers brought up were the need to layer annotations (annotation of annotations), the need to model bits of inline text, and the need for “sets” or aggregations of annotations (ORE Aggregation has been suggested as a possible method for this). The CODA implementers also found that the OA Data Model might be too heavy weight for the majority of simple annotation use cases.

3. **An early prototype for surfacing many annotations on arbitrary documents**¹²

Whaley presented the annotation work taking place at Hypothes.is. Hypothes.is is trying to build a web credibility platform, leveraging pre-existing work by the Open Knowledge Foundation¹³. Their key use case that they are trying to resolve is how to create anchors in documents such that the anchors survive the changes to the document. They also want to provide the ability for threading of annotations, such as with discourse annotations, in a similar manner to Wikipedia change requests (this use case also comes up for curatorial annotation use cases, especially in museums and similar venues). They also are concerned with quality control issues on annotations, including the general provenance issues of editing or deleting annotations.

4. **Annotation Supporting Collaborative Development of Scholarly Editions**¹⁴

Gerber presented the annotation work at Queensland. The team there has implemented an annotation repository and a set of annotation tools to aid scholars in the creation of discourse annotations. One of their key use cases is the creation of annotations that capture the variations between editions of manuscripts.

Queensland implementers also faced the issue of annotation sub-classes proliferating wildly. To reduce the overall number of annotation types they decided to make a distinction between annotating (which targets an object) and referencing (which does not explicitly target an object). In the case of a ‘See Also’ reference, they would model the reference as a link to the annotation body using a dc:relation rather than as a separate annotation.

---

¹³ [http://okfn.org/](http://okfn.org/)
Several issues brought up by the presentation included:

- Is it possible to distinguish the recurring variations case from the annotation of annotations; it was found that it isn’t possible to distinguish them.
- The presentation raised the question of how to draw the boundary between annotation typing and semantic tagging. The general consensus was that this should be left to individual communities to decide.
- Finally, Queensland provided their rule of thumb for typing annotations vs. using semantic tags. Generally if the information being provided by the typing is about the annotates relationship then a sub-class should be used, otherwise a semantic tag should be used.

5. **Developing an Annotation Framework for Fedora using OAC**\(^\text{15}\)

Ashton presented the annotation work at Brown. In Brown’s case they have a corpus of pre-existing annotations from a variety of annotation applications that have been used in the past. They needed an annotation model that would help them develop a system that facilitates migration of their corpus from multiple annotation data formats into a common format. The main issues raised by Ashton with regards to the OA Data Model were:

- The need for a fragment syntax identification
- UUIDs are not something native to systems like Fedora (or Drupal), so best practices for implementing them in those systems need to be developed.

6. **Automated Annotation of Biomedical Literature**\(^\text{16}\)

Livingston presented the annotation work at Colorado. Of all of the projects, the Colorado project makes the most unique implementation of the data model. Primarily a text mining use case, Colorado implementers are using the OA data model for machine-generated annotations of parts of speech and ontological concepts in a large corpus of biomedical journal articles. As such they need to build annotations that capture complex content and provide hooks for machine reasoning algorithms.

Graph annotations are a common type of annotation that emerges from their use cases. A novel extension that they have made to the data model is the ‘BasedOn’ relationship that explicitly feeds machine algorithms reference information about how higher annotation layers are composed from and relate to lower annotation layers. For example Anno3 makes the statement that Target1 is a sentence. It has BasedOn relationships with Anno1 and Anno2, which state that Target1a is a sentence part – subject and Target1b is a sentence part – predicate respectively. In this way the machine has a chain of evidentiary annotations displaying how conclusions are derived from more granular data. A key issue for the reasoning algorithms is their inability to cope with the ambiguity of the multiple targets case.


\(^\text{16}\) [http://www.openannotation.org/documents/colorado-oac-2012-07-v03.pdf](http://www.openannotation.org/documents/colorado-oac-2012-07-v03.pdf)
7. *Annotation of Digital Emblematica*\(^17\)

Cole presented the Emblematica annotation work at Illinois. The use cases Illinois is dealing with are primarily discourse annotation cases but, annotation typing was a major issue because the Emblem scholars had a tendency to type their annotations manually. In addition to annotation typing, the granularity of both bodies and targets, as well as developing or adopting an ontology more specific to discourse annotations were important issues.

8. *MapHub: A Demonstrator for Historic Map Annotations*\(^18\)

Haslhofer presented annotation work at Cornell. The Cornell project took maps from a variety of sources, including the Library of Congress, and built an annotation service layer over them that permitted users to annotate them. In addition to adding text, the prototype system leveraged an external look-up service to suggest tags for users to add to their annotations. The major observation coming out of the Cornell project was that the data model was too complex and made serialization of the annotations difficult.

9. *Annotation of Subscription Streaming Video Content*\(^19\)

Smith presented on the annotation work that was carried out at MITH and Alexander Street Press. For their experiment, a set of video annotation tools were built that could still provide access to the annotations even when the video content being annotated was secured behind a subscription wall. Major issues raised by their work included a question of how to use the model to express scaling of the target area communicated in the Selector (as overall scaling is communicated by the State).

10. *Annotation Middleware for Scholarly Publications and Resources*\(^20\)

Millman and Flores presented the annotation experiments at New York University. While they had a large series of experiments running with the data model, many of the results were still in the preliminary stages. Among the experiments they are running are scholarly discourse annotations via courseware, geo-annotation of maps, named entity annotation of digitized manuscripts, and e-book annotations. They currently have implemented a php client that can export OA-compliant annotations as JSON. They have plans to implement an RDF serializer in the future. The major issues NYU confronted when implementing the OAC data model revolve around provenance. Specifically they had unanswered questions regarding best practices for:

- Can annotations be deleted or edited?
- Should there be some form of versioning on the annotations?
- Should there be visibility controls on annotations?


Conclusions and Next Steps

After the presentations were completed Ciccarese discussed the overall plans for the W3C Open Annotation Community Group (next meeting scheduled for Sept. 18-19, 2012). Afterwards Cole reviewed the major themes that emerged over the course of the meeting and solicited advice on future community engagement initiatives to make in the future.

Attendees suggested that the NLP (Natural Language Processing) and Biomedical Informatics communities could be pursued but that the former (especially) would need working implementations to hook into as they were greatly affected by institutional inertia.

Cole assigned some overall action items to the group. These included:
- Sending budget updates to Jett.
- Sending individual project reports to Jett.
- Submitting final project invoices by 31 October [2012].
- Using disseminations on project results (e.g. conference papers, etc.) to recruit new contacts in additional communities that are potential stakeholders.

Finally, it was agreed that a series of brief white papers be written describing the issues exposed during the course of the meeting. Volunteers were asked to take the lead on these papers. The following list matches topics and volunteers:
- Semantic Tags vs. Multiple Bodies issue [Gerber]
- Ambiguity of Classes & Properties issue [Livingston]
- Body Content (String Literals / oax:hasSimpleBody) issue [Haslhofer]
- Annotation of Embedded (In-line) Resources issue [Smith]
- Annotation Versioning issue [Ciccarese & Leeds (Hypothes.is)]
- Multiple Targets issue [Brugman (related sequencing issue), Ciccarese & Livingston]
- BasedOn Property issue [Livingston]
- Selector Scaling issue [Cole & Smith]
- oa:hasStyle issue [Gerber & Sanderson]
- Class Sub-typing issue [Brugman]