A Publication Platform for Open Annotations

Hennie Brugman
Meertens Institute
P.O. Box 94264
1090 GG Amsterdam
E-mail: hennie.brugman@meertens.knaw.nl

Abstract

The Open Annotation Consortium introduced a generic model for representing annotations of resources and resource segments that complies with principles of the World Wide Web and Linked Data. This paper introduces a platform for storing, retrieving, searching, exchanging, harvesting and publishing Open Annotations on the web. It describes design considerations, functionality and architecture. Our Open Annotation server platform is set up as a distributed system with server instances that can exchange annotations in a peer-to-peer way. Each instance can persistently publish annotations using principles of the web and thereby adds 'annotatability' to annotations themselves and to annotation 'Bodies'. Additionally, the annotation platform provides efficient search and implements a Dashboard for server management tasks.
The web-oriented nature of the platform raises a number of interesting issues and opportunities that are discussed in some depth. For example, in general uploaded annotations do not have resolvable http URIs. Assigning those in not trivial. Indexing strategy, determining the boundaries of an annotation in an RDF graph and searching for annotations whose Body is somewhere else on the web are other issues that are discussed.

1 Introduction

Over the last decade a lot of progress has been made on the standardization and application of annotations for linguistic and multimodal resources. A family of linguistic annotation models and formats emerged that were based on graphs (Annotation Graphs, ATLAS, EAF, ANVIL, Exmeralda, etc). Discussions between the authors of these models led to some convergence and harmonization of models and formats, and also contributed to standardization processes as took place in ISO TC37/SC4 and resulted in standards like LAF (Linguistic Annotation Framework) (Ide, 2007).
More recently, the emergence of Semantic Web and Linked Open Data led to explorations of annotation models that were more generic with respect to types of annotated resources, more semantically oriented and more web based. (Brugman, 2008). A recent development along these lines is Open Annotation, an effort by the Open Annotation Collaboration (OAC) (Sanderson, 2011a), which is currently taken up by the W3C Open Annotation Community Group.

The Meertens Institute currently leads two projects where a range of multimodal annotation cases plays an important role. CATCHPlus is a valorization project that is associated with the Dutch CATCH research programme. CATCH includes a number of application driven research projects at large Dutch cultural heritage institutions. CATCHPlus builds tools and services on basis of research prototypes and demonstrators from CATCH. CATCHPlus annotation cases include annotation of images, sound and video recordings, text and music.
CODA (CATCHPlus Open Document Annotation) is an OAC Phase II project with funding by the Andrew W. Mellon Foundation. It focuses on two cases of annotation of scanned handwritten documents.
For both projects, it is required to support annotation of annotations, and annotation of 'annotation values' (for example, semantically annotate entities in transcription texts that are associated with an image region).

Currently, we are in the process of building an OAC compliant Open Annotation repository Service for CATCHPlus and CODA (expected delivery date is April 2012). Such a service is an essential part of the digital infrastructure needed to store, retrieve, exchange, search, publish and otherwise exploit heterogeneous annotations of multimodal, web based resources.

This paper starts with providing the necessary background about Open Annotation and two of its application domains. Section 3 then describes our Open Annotation Server platform and some of its planned applications. Section 4 discusses a number of issues raised during design and implementation. It will especially focus on issues related to the web-based nature of both the annotation model and the repository service. We will end with general conclusions and some remarks about the systems potential in section 5.

2 Open Annotation

Open Annotation (Sanderson, 2011a) is a generic approach to modeling of annotations as web based documents that are associated with web resources using World Wide Web and Linked Data principles. Sharing of annotations across clients, servers and applications is one of its main objectives. It promotes the use of publish/subscribe mechanisms (but does not prescribe a specific protocol for that).

2.1 Open Annotation data model

An Annotation is defined as "a document containing references to the Body and Target, which the Body is somehow about". Annotations, Bodies and Targets are all resources that are identified with URIs, can all have additional metadata, and can each have a different author. See figure 1 for an illustration of the basic model, enriched with additional metadata and relations.

Open Annotation is generic in the sense that Bodies and Targets can be of any media type.

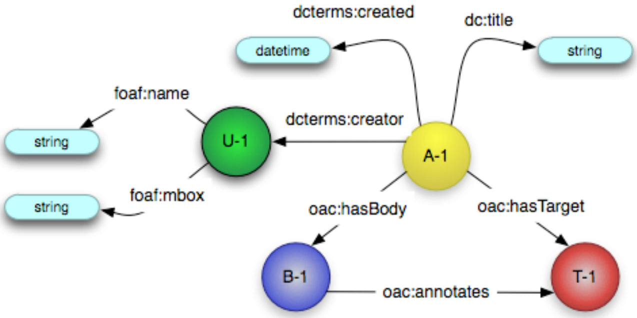

Figure 1: Open Annotation baseline model with additional properties and relations (copy from Open Annotation beta specification). A1 is represents an Annotation

Bodies and Targets may also be *parts* of resources. Open Annotation supports the use of fragment URIs (to address parts of X/HTML, PDF, plain text and XML), media fragment URIs (for spatial and temporal segments) and ConstrainedTargets, which is a generic way to represent diverse types of constraints, where interpretation of the constraints is up to the annotation client.

2.2 SharedCanvas

SharedCanvas (SC) (Sanderson, 2011b, 2011c) is an application of Open Annotation to the domain of scanned (scholarly) documents. It is an extension and specialization of the OAC model. The basis is a two-dimensional Canvas abstraction that can be annotated with ImageAnnotations on the one hand and TextAnnotations on the other hand. TextAnnotation and ImageAnnotation are specializations of the OAC 'Annotation' class. SharedCanvas adds constructs for groups of Canvases (Sequence, Range) and (ordered) groups of Annotations that are implemented on basis of OAI-ORE aggregations (see W3C specification).

The two cases from the CODA project explore the SharedCanvas extension of Open Annotation. This implies that the generic Open Annotation repository Service that is being build has to be able to deal with SharedCanvas data in a graceful way. It has to be able to store and retrieve SC data in its entirety. Furthermore, queries that are motivated by CODA use cases have to be handled by the service as adequately as possible. These may include queries for additional SC classes and properties or queries dealing with the two-dimensional spatial structure of the annotation Targets.

2.3 Annotations of linear data

A number of CATCHPlus and CODA cases are dealing with annotations of one-dimensional data, such as audio, video, music and text (there is also another OAC phase II project that deals with applying Open Annotation to streaming video). Typical for annotation of one-dimensional data is the occurrence of several, sometimes complexly interrelated layers of annotations. For this type of annotation, relevant queries include queries about overlapping and sequencing.

Again, our Open Annotation repository Service has to be able to deal gracefully with such specializations.

3 Open Annotation Server platform

Open Annotation presents a data model, and explicitly states that "no client-server protocol for publishing /updating/deleting annotations will be specified. Rather, the specifications will take a perspective whereby clients publish annotations to the Web and make them discoverable using common Web approaches. Such an approach does not require a preferred annotation server for a client, yet it does not preclude one either." (See: OAC Guiding Principles).

However, almost all of the currently existing tools and services that produce annotations have no facilities to persistently publish them to the web according to World Wide Web and Linked Data principles, nor do they have ways to maintain them or make them efficiently searchable once they are published.

We are currently implementing such a publishing facility that can be used in conjunction with existing annotation production and exploitation systems. This chapter describes the system, its design considerations and planned applications.

3.1 Vision and functionality

The Open Annotation Service platform we are currently building is not envisioned to be a single centralized 'data silo'. Rather, it will be an easily installable and configurable web service component that can be applied by individual users, by working groups, by institutions or even by consortia. This implies that the system has to scale very well.

For efficient collection of sets of annotations in order to be able to search efficiently, the system has built-in OAI-PMH data providers and harvesters. Running instances of the service can harvest sets of annotations from each other in a 'peer to peer' manner and index them.

Annotations can be stored in one of the repository instances by means of an *SRU/Update* interface. This can be done either one by one or batch wise, using the Open Annotation RDF/XML serialization as it is published on the Open Annotation website as input format.

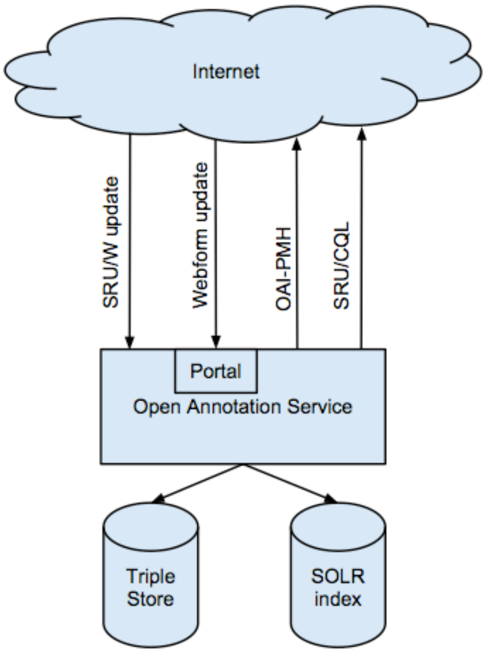

Figure 2: Open Annotation Service architecture

Annotations as annotation tools or services produce them typically have no stable, resolvable http URIs of their own, which is a requirement for persistent publication. The repository system therefore assigns these URLs when necessary, while maintaining original identifiers where these exist. For a discussion of this handling of URIs see section 4.1. Note that in principle it is also possible to use actionable persistent identifiers (e.g. actionable Handles) as URLs for annotations.

Many of the standard fields/properties of imported or harvested Open Annotation data will be indexed for efficient searching. This searching is currently done using SRU/CQL. At a later stage we intend to add a RESTful search API that will be discussed and if possible harmonized with partners in the Open Annotation Collaboration.

The Open Annotation Service platform will as a baseline implement queries for constructs of the core OAC annotation model. We foresee the need for queries for generic constructs that are not part of the core OAC model, such as for *groups* of annotations. In collaboration with OAC we are contributing to a recommendation document that will be published together with the OAC specification and that specifies how best to deal with constructs like grouping. The service will support searches for these recommended grouping constructs as well. Finally, specialized queries for specifics of SharedCanvas annotations or annotations of linear data are not supported. However, the system will be implemented in a modular way that allows easy extension with specialized query interfaces.

In principle, instances of the Open Annotation Service will be open for reading and searching operations. Updating the information in the repository will be protected by an API key to be used with the SRU/Update interface.

Each instance of the service will have an interactive Dashboard, that allows authorized users to perform management of API keys, to configure and schedule OAI-PMH harvest jobs and to do simple interactive searches on the repository's annotation content.

3.2 Architecture

Figure 2 shows the architecture of an instance of the Open Annotation repository service. It is mainly constructed on basis of existing components from the Meresco open component library (http://meresco.org) that are widely applied, well tested and available under open source license. Meresco contains off-the-shelf components for SRU/W Update, OAI-PMH, SRU/CQL and many other components that are related to metadata management and search. For the internals of our service platform, both a triple store (OWLIM) and Apache Solr are used.

Among the project specific additions are modules for custom indexing, a built-in URN resolver and processing logic for assignment of identifiers.

Finally, an interactive Dashboard is part of the system.

3.3 Applications

We have a number of concrete applications for the Annotation repository Service.

- Store and publish manual annotations for scans of index books of the Queen's Cabinet, a collection provided by the Dutch National Archive.
- Text from the Bodies of these manual annotations will be processed by a Named Entity Recognizer web service. This service will generate an additional 'layer' of annotations that will also be stored on the Annotation server.
- Store and publish manual transcriptions for scanned images from the Sailing Letters collection. This collection is also provided by the National Archive. Transcriptions are created by volunteers in a project run at the Meertens Institute.
- As an aid for manual transcription of scanned images we will create a service that automatically detects bounding boxes around written lines in a scanned image. Both input and output for this service will contain annotations that are in Open Annotation format. Results can be stored on an instance of the server and later reused for manual transcription.
- CATCHPlus has a number of sub-projects that deal with annotations. Included are annotations of text, images, speech data, video and music.

4 Discussion

In this section we present a more in-depth discussion of some of the non-trivial design decisions that we had to make and that mostly have to do with the web-based nature of both Open Annotation and our Open Annotation publishing platform.

4.1 Handling of identifiers

Annotations that are uploaded to the system are supposed to be represented using RDF/XML. In practice, there is

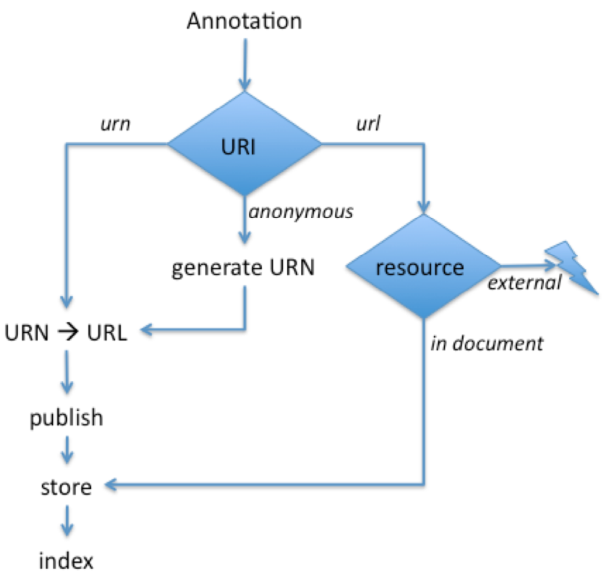

Figure 3: Flow chart illustrating how identifiers of incoming annotations are processed

some variety in RDF conventions used.

Also, according to the Open Annotation specification URIs used for Annotations, Bodies, Targets or creators can take the form of URNs (used for internal references) or URLs (used as external, resolvable references). Annotation production systems (such as tools for manual annotation) typically do not produce annotations that have a web presence. In such cases annotations are identified by means of URNs or not at all.

One of the functions of the Open Annotation repository service is that it publishes annotations. Furthermore, our use cases require that Annotations are 'annotatable' web documents by themselves. Therefore, the system assigns URLs where they are necessary but missing.

All of this implies that processing identifiers in the incoming annotation data is not trivial. Figure 3 shows a flowchart that illustrates how URIs of incoming Annotations are processed. Annotations identified by URLs are either directly stored and indexed, or not processed, because they are references to Annotations that are defined and published elsewhere. For Annotations that do not have a URL the system creates one, based on a URN. The original URN, if present, is stored for reference (in a dc:identifier field). Such annotations can be retrieved from the system by either their URN or URL.

Similar and sometimes even more complex flows are implemented for processing of identifiers for Annotation Bodies and authors.

4.2 Import and indexing strategy

For efficient searching it is necessary to create indexes for all fields that we want to search on. Imported annotations can turn out to be embedded in richer RDF graphs (e.g. SharedCanvas documents). Since our system is essentially an Annotation server we chose to start processing RDF data at the Annotation objects and only process what is connected to these annotations, effectively ignoring all other RDF statements. This may for example have consequences for the 'searchability' of SharedCanvas data, since not all of the SharedCanvas data will be imported in this way.

Since we at least want to be able to search on all attributes of the OAC core model, all properties shown in figure 1 are indexed. One important query is for all Annotations that are associated with any segment of a given Target. To support such queries Target URLs are indexed both with and without additional fragment identifiers.

Users may want to search on information that is only referred to by URLs pointing from the annotations to external resources. For example, authors may be identified by an external foaf profile or annotation Bodies are represented by URLs pointing to resources elsewhere on the web. To some extend the Open Annotation Service can resolve such URLs and embed and index the retrieved information. For example, the system tries to retrieve foaf:name and foaf:mbox for a 'creator' URL and index those to be able to search on creator information.

For some textual types of Body resources we also follow this strategy. For such Bodies it will be possible to search on text content even though this text is not included in the Annotations inside the repository.

4.3 Annotation boundaries

An Open Annotation is represented by a graph that consists of a set of triples. This leaves ambiguity to what it means to 'replace' or 'delete' an annotation in the repository. Either such operations are performed only on the relevant triples, leaving other triples untouched, or they involve a complete annotation graph delimited by a well-defined *annotation boundary*.

Also, we have to define the boundaries of what sub graphs to return when someone queries for annotations. Which properties do we take into account? How many levels of triples do we consider to be part of the graph belonging to some annotation?

We currently see two alternative approaches. Either we use conventions at the RDF level, such as Concise Bounded Descriptions (CBD)[1] or we define our own heuristics based on specific vocabularies used for properties in the annotation data (OAC, OAI-ORE, DC, other). We tend to use the latter approach.

4.4 Other considerations

This section raises a couple of issues of diverse nature.

- An optional extension that is seriously considered is to store and publish *annotation schemes* in the repository as well. Annotation schemes define templates that can be used to configure annotation production systems or query formulation systems. The Open

[1] http://www.w3.org/Submission/CBD/

Annotation data model can be extended with sub types of Annotations. Annotation schemes can be defined on basis of such sub types.

- The Open Annotation repository stores objects of type Annotation as well as sub types of it. This *type-inheritance* can be exploited by the query interface. It is possible to search for Annotations both at a generic level and as instances of specific sub types.
- Incremental OAI-PMH harvesting requires that the system maintains *date-time stamps* for last modification times of all of its harvestable objects. Besides that, the system will be able to search for annotations that are time-dependent, in the sense that they refer to representations of resources at a specific time. The core Open Annotation model explicitly supports time-dependent annotations.
- The Open Annotation Service is only one of the deliverables of the CATCHPlus project. Another product is OpenSKOS, a web service based publication and search platform for vocabulary data that can be mapped to the W3C SKOS model (Miles, 2009). Open Annotations can refer to vocabulary concepts in OpenSKOS by URL.

5 Conclusions

Work on annotations and annotation repository systems has been going on for quite some time. However, an annotation model that is based on principles of the World Wide Web and Linked Data, used in combination with an annotation repository service that enables publishing of annotations on the web raises a number of new issues to be tackled.

The Open Annotation Service implements possible strategies for indexing annotation information and for assignment of URLs to annotation resources. Since the service also publishes URLs for annotation Bodies, these Bodies can be subject to further annotation. This enables incremental enrichment of web resources by adding layers of annotations on top of other layers.

The Annotation repository service is not designed as a single, central data silo, but is meant to be used as a system with multiple running instances that can be easily deployed at different sites of different scale. Users or user communities can collect published sets of annotations from several service instances for specific projects or use cases and make these locally searchable. This provides new opportunities for collaboration.

For example, in projects like CATCH and CATCHPlus cultural heritage institutions can combine heterogeneous annotation data for their online collections in one shared annotation repository. Searching or browsing the annotation repository then gives direct online access to relevant segments of these online resources (relevant text paragraphs, image regions or video scenes, etc).

The Open Annotation Service will be available as easily installable package. Source code will be made available on the web under open source license.

6 Acknowledgements

We would like to thank our collaborators at the OAC for their support and feedback. We also thank the funders of CATCHPlus: the Netherlands Organisation for Scientific Research (NWO), and the Dutch ministries for Education and Economic Affairs. Finally, we would like to thank the Andrew W. Mellon foundation for funding our CODA Open Annotation phase II sub-project.

Special thanks goes to Erik Groeneveld and Johan Jonkers from Seecr. They raised and helped solve many of the issues discussed in this paper.

7 References

Brugman, H. Malaise, V. Hollink, L. (2008). A Common Multimedia Annotation Framework for Cross Linking Cultural Heritage Digital Collections, In *Procs of 6th International Conference of Language Resources and Evaluation*, Marrakech, Morocco, May 2008

Ide, N., Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (Eds.), *Evaluation of Text and Speech Systems*, Springer, 263-84.

Miles, A., Bechhofer, S. (2009). SKOS Simple Knowledge Organisation System Reference. *W3C Recommendation* 18 August 2009.

Sanderson, R., Van De Sompel, H. (2011a). Open Annotation. Beta Data Model Guide. http://www.openannotation.org/spec/. 10 August 2011.

Sanderson, R. et al. (2011b). SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination, In *Procs of 11th Joint Conference of Digital Libraries*, Ottawa, Canada http://www.arxiv.org/abs/1104.2925