Annotopia: An Open Source Universal Annotation Server for Biomedical Research

Paolo Ciccarese and Tim Clark

Massachusetts General Hospital and Harvard Medical School, Boston MA

paolo.ciccarese@gmail.com; tim_clark@harvard.edu

Abstract

Annotopia is an open source, open services platform for creating, managing, manipulating and sharing open annotation using the W3C Open Annotation Data Model. It can create and/or manage annotation of HTML, PDF, and other resources including data and ontology concepts, with text, semantic tags, and other annotation types. It supports fine-grained permissions on annotations. Annotopia is a Swiss-army knife for W3C Open Annotation system developers, eliminating many otherwise challenging backend development tasks.

Keywords: annotation, biomedical, entity recognition, semantic web

1 Background

Annotation of documents and databases on the Web is a core aspect of the Web's interactivity [1], but until recently, annotations have been second-class objects, tied permanently to the applications and servers that host them [2]. This is a significant missing feature for the scientific and biomedical community, which increasingly relies on the Web as its primary means of knowledge dissemination and group interaction.

As a result of this feature gap, comments, discussions, semantic tags, references, and other annotations on biomedical publications, are atomized across disparate servers and media type based representations. Our goal is to fill this gap, making annotations first-class, independently-managed objects on the web.

Projects from distributed hypermedia research programs in the 1980's, upon which many aspects of the early Web were based, actually had several of these properties [3-5]. Berners-Lee's inspired stripping down of these models into "the simplest thing that could possibly work" [6], laid the basis for a transformative global, collaborative development of the Web and its technologies [7], but necessarily removed such features.

In the early 2000's, W3C's Annotea project began to attempt restoration of some of the lost features based on the modern web architecture [8]. Annotea was a foundation for later annotation models and systems focused on biomedical annotation [9, 10]. Similar models were developed for digital humanities use cases. These specifica-

tions were merged to develop a more diverse, community-based specification, the W3C Open Annotation Data Model [11], now on standards track in the W3C.

While various annotation tools are now available and in use, current annotation platforms use different representation formats. Such tools normally provide little or no means to export the annotation in a usable or reusable fashion. The Open Annotation model is directed at solving the format interoperability issue, but interoperability on a large scale has not been showcased yet. It requires the existence of special tooling to handle storage and distributed integration of the Open Annotation Data Model (OADM) format annotation, and our analysis indicated that this is a server-side issue.

Furthermore, updating existing software is never an easy task and creating new software with an Open Annotation backbone requires significant knowledge of the OADM specifications and introduces software constraints. It appears that most annotation efforts focus a lot of their energy on the front-end or client as user interaction is key for adoption and the technical difficulties to work with different operative systems and browsers are not trivial.

Moreover, the annotation projects of which we are aware, all rely on different back-end software. We argue that developing many different back-ends, which perform very similar operations, results in higher community costs and in a slower penetration of the Open Annotation specification. OADM services, if not based on a common service model, will need to be implemented in several pieces of software before having different systems communicating and exchanging content.

Lastly, we also have noted that the existing annotation back-ends implement very similar features that could be coded once and easily serve multiple clients. Within the scope of an extensible architecture, services like: (i) Open Annotation compliant storage, (ii) text mining, (iii) entity recognition, (iv) image analysis and (v) Linked Data mashups could be implemented once as common services. This approach could reduce the development time and the cost of future annotation platform whose developers will be able to focus on new features and components without the necessity of reinventing the common functionality.

2 Methods

Our group developed, and will demonstrate in this workshop, Annotopia: the first W3C OADM-compliant, biomedically-oriented Open Annotation server, in response to these challenges. Annotopia is a joint project of researchers at the Massachusetts General Hospital and Eli Lilly & Company. It is an open-source product (https://github.com/Annotopia). Annotopia operates as a Swiss Army Knife for annotation. It facilitates creation of interoperable annotation platforms, by providing an extensible back-end solution supporting the open W3C standard. Thus it allows developers to focus effort on client software, reducing development time and resources.

Annotopia is constructed so that every Annotopia instance can support integration with (i) multiple annotation clients (ii) other Annotopia servers (iii) other Open Annotation compliant servers (iv) other non Open Annotation compliant servers (v) existing text mining services (vi) pre-computed text mining results (vii) ontology ma-

nagement platforms and custom databases for generating structured annotation (viii) Linked Data SPARQL end-points and much more.

Annotopia incorporates features and ideas from two other annotation servers we developed in the recent period: CATCH and Domeo. We have extensively described Domeo elsewhere [10]. CATCH supports HarvardX Massively Open Online Courses (MOOCs), with textual and video annotation, for classes as large as > 20,000 students.

While CATCH and Domeo focus on annotation of video, images and textual documents (HTML and PDF), Annotopia allows in addition, annotation of data, or of anything that is uniquely identifiable, even concepts in ontologies.

Annotopia has been integrated and tested against the Annotator.js client, the Domeo Web annotation toolkit, and the Utopia PDF viewer [12].

3 Annotopia architecture and technologies

Annotopia consists of a modular architecture providing a series of extension points. Extension points are necessary for handling custom structured annotation types as well as an always-increasing amount of external services to be integrated with the platform through appropriate connectors. The core platform is written in Java/Groovy [13] making use of the Grails [14] web application framework. The Grails plugin infrastructure has been extensively exploited for realizing the modular approach.

A high level view of the architecture is shown in Figure 1.

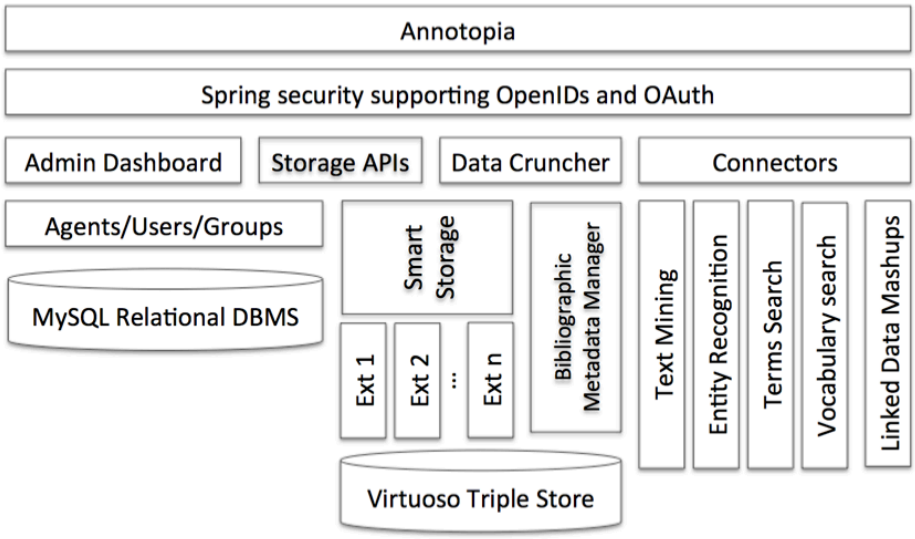

Fig. 1. - The high-level Annotopia architecture. Each block corresponds approximately to a software plugin or module.

4 Demonstration

Our demonstration will showcase the annotation storage, search, and textmining integration capabilities of Annotopia. We will also demonstrate interoperability between multiple HTML and PDF article representations. We expect in future to be able to demonstrate direct database annotation as well.

5 References

1. O'Reilly T: What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. In.: O'Reilly Network; 2005 [http://www.oreillynet.com/lpt/a/6228].
2. Ciccarese P, Soiland-Reyes S, Clark T: Web Annotation as a First-Class Object. IEEE Internet Computing 2013, Nov/Dec 2013:71.75
3. Bechhofer S, Goble C: COHSE: Conceptual Open Hypermedia Service. In: Annotation for the Semantic Web. Edited by Handschuh S, Staab S. Amsterdam: IOS Press; 2003.
4. Carr L, De Roure D, Hall W, Hill G: The Distributed Link Service: A Tool for Publishers, Authors and Readers. In: Fourth International World Wide Web Conference: December 11-14, 1995; Boston, Massachusetts, USA. World Wide Web Consortium (W3C) 1995:
5. De Roure D, Carr L, Hall W, Hill G: Enhancing the Distributed Link Service for multimedia and collaboration. In: Distributed Computing Systems, 1997, Proceedings of the Sixth IEEE Computer Society Workshop on Future Trends of: 29-31 Oct 1997 1997. 330-335
6. Venners B: The Simplest Thing that Could Possibly Work: a Conversation with Ward Cunningham, Part V. Artima Developer 2004 [http://www.artima.com/intv/simplest.html].
7. Jacobs I, Walsh N: Architecture of the World Wide Web, Volume One. In: W3C Recommendation World Wide Web Consortium; 2004 [http://www.w3.org/TR/2004/REC-webarch-20041215/].
8. Kahan J, Koivunen M-R, Prud'Hommeaux E, Swick RR: Annotea: An Open RDF Infrastructure for Shared Web Annotations. In: WWW10 International Conference: May 2001 2001; Hong Kong. World Wide Web Consortium: [http://www10.org/cdrom/papers/488/index.html].
9. Ciccarese P, Ocana M, Castro L, Das S, Clark T: An open annotation ontology for science on web 3.0. J Biomed Semantics 2011, 2(Suppl 2):S4
10. Ciccarese P, Ocana M, Clark T: Open Semantic Annotation of Scientific Publications with DOMEO. Journal of Biomedical Semantics 2012, 3(Suppl 1):S1 [http://www.jbiomedsem.com/content/3/S1/S1].
11. Sanderson R, Ciccarese P, Sompel HVd, Bradshaw S, Brickley D, Castro LJG, Clark T, Cole T, Desenne P, Gerber A, Isaac A, Jett J, Habing T, Haslhofer B, Hellmann S, Hunter J, Leeds R, Magliozzi A, Morris B, Morris P, Ossenbruggen Jv, Soiland-Reyes S, Smith J, Whaley D: W3C Open Annotation Data Model, Community Draft, 08 February 2013. W3C 2013 [http://www.openannotation.org/spec/core/].
12. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D: Utopia documents: linking scholarly literature with research data. Bioinformatics 2010, 26(18):i568-i574 [http://bioinformatics.oxfordjournals.org/content/26/18/i568.abstract].
13. Henry K: A crash overview of groovy. Crossroads 2006, 12(3)
14. Rocher G, Brown J: The Definitive Guide to GRAILS. Berkeley CA: Apress; 2009.