Targeted Annotation of Scientific Literature and Data Resources in Invenio Digital Libraries

Adrian-Tudor Pănescu[1], Tibor Šimko*[1], and Christine Vanoirbeek[2]

[1]Department of Information Technology, CERN, Geneva, Switzerland
[2]MEDIA Research Group, EPFL, Lausanne, Switzerland

February 10, 2014

Abstract

We describe a new targeted annotation framework developed within the Invenio digital library platform in order to answer collaborative annotation needs of high-energy physics user community. After a brief illustration of the most common use case, we describe the new facility and explore its relation to W3C RDF and JSON-LD recommendations as well as to the W3C Open Annotations data model draft standard.

1 Use case

The experimental particle physics community, using CERN LHC accelerator to study the basic constituents of matter, consists of multi-national, multi-institute, multi-person collaborations. ALICE, ATLAS, CMS, and LHCb collaborations may comprise up to 3,000 physicists. The publishing workflows of these collaborations include multiple commenting rounds for paper drafts that are restricted within the collaboration before papers become public[1]. The commenting partially happens on an Invenio[2]-powered CERN document server[3].

The structured annotation habits of collaborations include referring to page, line, equation or figure in a free text from. The principal authors of the draft then need to aggregate and process comments left by collaboration members. This task is often non-trivial due to the sheer number of authors and the heterogeneous nature of commenting practices. As a concrete example, one of conference papers received more than thousand annotations in 180 comments coming from 50 persons during its pre-publication stage. (E.g. the most extensive comment contained more than 100 annotations targeted to various parts of the paper draft.)

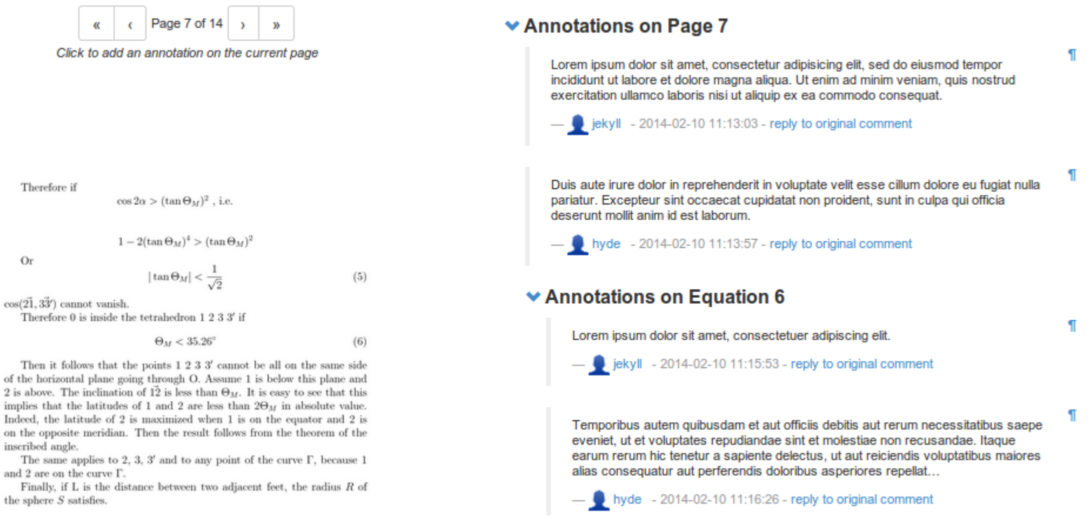
2 Targeted annotation system

In order to answer these targeted annotation needs, a new module was developed for the Invenio digital library platform. The module aims to provide an easy way to capture user annotations in a structured format so that future aggregations can be easily constructed.

When displaying a paper draft, a structured viewer permits to page through the document by showing a page preview on the left-hand-side, while annotations of users for the given page are show on the right-hand-side, see Figure 1.

Users can enter annotations via so-called “online” and “offline” modes. When online, user can click on a page which brings up a simple structured editor that permits to enter notes for the concrete logical page (P.7), section (S.7), paragraph (PP.7), line (L.7), figure (F.7), table (T.7), equation (E.7), or

*Corresponding author.



Therefore if

$$\cos 2\alpha > (\tan \Theta_M)^2 \text{ , i.e.}$$

$$1 - 2(\tan \Theta_M)^4 > (\tan \Theta_M)^2$$

Or

$$|\tan \Theta_M| < \frac{1}{\sqrt{2}} \quad (5)$$

$\cos(\vec{21}, \vec{33'})$ cannot vanish.

Therefore 0 is inside the tetrahedron 1 2 3 3′ if

$$\Theta_M < 35.26^\circ \quad (6)$$

Then it follows that the points 1 2 3 3′ cannot be all on the same side of the horizontal plane going through O. Assume 1 is below this plane and 2 is above. The inclination of $\vec{12}$ is less than Θ_M. It is easy to see that this implies that the latitudes of 1 and 2 are less than $2\Theta_M$ in absolute value. Indeed, the latitude of 2 is maximized when 1 is on the equator and 2 is on the opposite meridian. Then the result follows from the theorem of the inscribed angle.

The same applies to 2, 3, 3′ and to any point of the curve Γ, because 1 and 2 are on the curve Γ.

Finally, if L is the distance between two adjacent feet, the radius R of the sphere S satisfies.

Annotations on Page 7

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. ¶

— jekyll - 2014-02-10 11:13:03 - reply to original comment

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. ¶

— hyde - 2014-02-10 11:13:57 - reply to original comment

Annotations on Equation 6

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. ¶

— jekyll - 2014-02-10 11:15:53 - reply to original comment

Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet, ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat… ¶

— hyde - 2014-02-10 11:16:26 - reply to original comment

Figure 1: Example of document viewer page (lhs) with aggregated user annotations (rhs).

reference (R.[7]). The leading markers can be nested: for example, marker "P.2: T.3: L.4" refers to line four of the table three on the second page. When offline, users can prepare block of text using the same markers. The text will be then parsed into targeted annotations upon submission.

The framework also permits nested annotations, i.e. to reply to a comment, to annotate an annotation, etc, without any depth restriction.

Finally, the new annotation framework permits to annotate any URI. This makes it possible to use the same commenting and annotating facility to discuss upon any resource handled by the digital library or the application itself, e.g. simple discussion forum on help pages.

It should be noted that similar annotating frameworks exist, such as Annotator [4] or Pundit [5]. While these standalone solutions could cover some of our needs — especially URI-based annotations — our use case usually requires strict restriction of annotations within the given digital library platform. (E.g. collaboration notes on paper drafts remain private even after publication.) The present work therefore attempts at providing an organic evolution of existing commenting and tagging facilities of Invenio towards targeted, structured and linked annotations, all the while providing pluggable interfaces to existing external third-party tools and solutions.

3 JSON model and interoperability

The persistence of annotations is achieved by storing an appropriate JSON representation in a document database, namely MongoDB [6]. (Under way is the integration with the PostgreSQL relational database [7] that offers extensive JSON indexing and searching capabilities as of version 9.3.) The schematic JSON document model used for general URI annotations is depicted in Figure 2. The reasons for opting for a JSON representation are two-fold.

The first reason is related to the extendibility of the model; for example, in order to allow for annotations with attachments, a single field specifying the storage location of the files needs to be added to the model described in Figure 2. Coming back to the targeted record commenting use case presented in the previous section, such a specific annotation could be represented by using an extended version of the model, such as the example in Figure 3. (Only specific fields depicted.) All these different types of annotations can be stored homogeneously in a document database, while an SQL representation would likely require distinct tables for each specific annotation type.

The second reason for choosing a JSON representation is related to the possible dissemination of annotations outside the Invenio platform. The Open Annotation Core Data Model [8] is a inter-operable framework proposed by the World Wide Web Consortium (W3C) for creating, connecting, and sharing

```
{
  id: UUID, // annotation identifier
  who: Number,  // user identifier
  where: *, // annotation target (e.g. URI, record identifier)
  what: *, // the content and its properties
  when: Timestamp,  // annotation creation date and time
  perm: Object  // permissions e.g. public, private (group-restricted)
}
```

Figure 2: General JSON model for annotations.

```
{
  where: { URI: "http://cds.cern.ch/record/897121",
           target: "P.2: T.3: L.4" },
  what: { title: "Small correction",
          body: "There is a typo near the word xyzzy." },
  type: "grammatical"
}
```

Figure 3: Example of JSON fields specific for targeted annotations.

associations and annotations across the Web. It uses RDF [9] for modelling and the JSON-LD [10] format is recommended for serialisation. As JSON-LD is a JSON-based format specialized for linked data, this should allow us to easily export annotation data originating on an Invenio platform and use any of the available third-party tools for the semantic study of annotations.

4 Conclusions

A new targeted annotation module for Invenio digital library was developed in order to permit structured annotation of scientific literature, as motivated by the needs of high-energy physics experimental collaborations. The module permits “online” and “offline” ingestion of targeted annotations and displays an aggregation of notes around logical units of text and/or data. The captured information is designed to be exportable as JSON-LD or RDF XML.

References

[1] L. Marian, *The Workflow of LHC Papers*, Computing in High Energy and Nuclear Physics (CHEP), May 21–25 2012, New York, USA.

[2] Invenio digital library platform. `http://invenio-software.org/`

[3] CERN document server. `http://cds.cern.ch/`

[4] Open Knowledge Foundation, *Annotator*. `http://annotatorjs.org/`

[5] Net7 srl, *The Pundit*, 2013. `http://www.thepund.it/`

[6] MongoDB document database. `http://mongodb.org/`

[7] PostgreSQL relational database. `http://postgresql.org/`

[8] The World Wide Web Consortium (W3C), *Open Annotation Data Model*, Community Draft, February 2013. `http://www.openannotation.org/spec/core/`

[9] The World Wide Web Consortium (W3C) RDF Working Group, *Resource Description Framework (RDF)*, February 2004. `http://www.w3.org/RDF/`

[10] The World Wide Web Consortium (W3C), *JSON-LD 1.0: A JSON-based Serialization for Linked Data*, Recommendation, January 2014. `http://www.w3.org/TR/json-ld-syntax/`