Common Annotation, Tagging, and Citation API

Philip Desenne, Academic Technologies Senior Product Manager, FAS ATG
Paolo Ciccarese, PhD, Instructor in Neurology and Senior Software Engineer, HMS
Martin Schreiner, Head of Maps, Media, Data and Government Information, HCL

The Project Idea

We here propose to create a prototype for a Common Annotation,Tagging and Citation (CATC) API . CATC is a unified public open API (Application Programming Interface) that will enable storing, searching, discovering, sharing and analyzing scholarly annotations produced on four digital media types - text, image, audio and video - across existing pedagogical and research tools at Harvard. Many of these tools already allow basic forms of annotation, however, they are using different formats and storage systems. These de-facto information silos are limiting knowledge sharing and therefore obstructing annotation aggregation and analysis. The CATC overcomes these limitations by providing an independent and shared hub for annotation and by promoting interoperability through the use of the Open Annotation Model produced by the W3C Open Annotation Community group (http://www.w3.org/community/openannotation/) co-chaired by Dr. Ciccarese.

What problem it will solve

In the current digital paradigm for research and teaching, there is an enormous and growing need to track, search and discover an overwhelming number of relevant digital resources, scholarly annotations and citations. With academic work across all disciplines now requiring detailed engagement with all forms of media including images, video, audio, and hypertext; a service for easily citing, tracking, searching and discovering relevant resources and annotations across all of these forms is paramount.

With search and discovery being central to this paradigm, the library's role as builder and provider of access to both high quality repositories and high quality standardized metadata are key elements in creating a global research environment. The CATC API is a means for the library to extend its online catalogs and repositories more directly into scholarly work of researchers and the learning activities of students. Conversely, it is also a means for researchers and students to more easily and dynamically connect their work to the library's repositories and catalogs in a highly collaborative way that increases discoverability.

Since antiquity scholars studied manuscripts or printed text by marking passages by hand, placing notes, symbols, underlines and highlights on documents. Time-honored handwritten marginalia has persisted through millennia and is still accessible today for analysis by scholars.

With the advent of digital text, and the rapid emergence of primary scholarly sources in other digital media formats, such as images, audio and video, the traditional annotation model has shifted into a new digital realm. But current tools that operate in this digital realm are lacking interoperable standards that will guarantee the longevity of the new digital marginalia.

In fact, many clients exist to create annotations, but in most cases the annotations are kept in privately owned, difficult-to-access digital silos and are normally structured in proprietary metadata formats. Moreover, annotation content in these proprietary systems is ephemeral, and usually tends to disappear or is no longer accessible when a tool has reached the end of its digital lifespan. The CATC API would provide an open access scholarly environment in which search, discovery, collaboration, permanence and provenance are enhanced.

Method to solve the problem

We propose to develop the CATC, an annotation gateway service that will:

- Enable interoperability of existing pedagogical and research tools at Harvard that already provide annotation and citations capabilities.
- Realize the annotation exchange by promoting import and export of annotations using the Open Annotation Model produced by the W3C Open Annotation Community group (http://www.w3.org/community/openannotation/) co-chaired by Dr. Paolo Ciccarese.
- Allow encoding the annotation content in several formats such as: Resource Description Framework (RDF), structured text formats such as Text Encoding Initiative (TEI) and simple plain text.
- Offer methods to search annotations by types, authors, ontological terms or keywords.
- Streamline the aggregation of cross-referenced annotations for custom comprehensive thematic analysis, data/knowledge mining and visualizations

Connection with existing activities

A growing proportion of academic teaching, learning and research is wrapped up in digital authoring and manipulation of library collections across all media for student assignments, academic presentations, lecture videos and scholarly publication. In this current paradigm for academic work our prototype would accomplish two essential things: 1. a demonstration of possibilities for more easily integrating the library's resources with the personal research materials of our users and 2. the enhancement of discoverability with the scholarly work of others. In essence, the CATC API would provide the necessary virtual mortar in connecting the building blocks of our virtual library space for a new level of access and discovery in and through our collections and online catalogs. These building blocks are the tools currently being developed and evolving through our Library Lab and HILT projects.

The Common Annotation,Tagging and Citation will be relying on persistent digitally archived resources such as the ones currently available through the new Harvard Library Digital Repository Service (DRS 1 and 2).

The CATC will provide a valuable API for Library Lab projects or Harvard Initiative for Teaching and Leaning (HILT) projects, online courses and research projects that rely on some form of annotation (tagging, commentaries, citations, etc) or contemplate the use of cross-referenced annotations.

The following Library Lab Projects could benefit from the CATC API (we have had conversations with several of the recipients that have expressed interest in the API):

- A Reusable Tablet-Based Application for Library Collections
- Connected Scholar
- Highbrow: A Textual Annotation Browser
- Zone 1
- Social Tagging for Archival Collections
- Zeega
- Transcription for Improved Research, Teaching and Learning at Harvard

HILT Projects that could benefit from the CATC API:

- Focus on Teaching: A collaborative venture to develop pedagogical insights, ambitions, and techniques
- The Connected Scholar (Phase 2)
- The lecture in 21st century learning: Reconstructing and revaluing our oldest teaching asset
- Development of a multimedia textbook
- Enhancing learning through hands-on exploration in a dynamic cross-disciplinary geospatial web platform

In addition to these projects, several courses within FAS already rely on annotations for teaching and learning. Currently Harvard iSites uses two tools that contain annotation features: The Collaborative Annotation Tool and the Video Production Tool. We would like for these tools to interoperate through Open Annotation standards by making use of the API.

Examples for Interfacing CATC API with other Library Lab and HILT projects
Projects such as *Highbrow: A Textual Annotation Browser*, rely on annotation data that is harvested or created through proprietary methods that are not fully standardized and are labor intensive to obtain (HTML page scraping is one example). The CATC API would bridge the annotation data acquisition gap for Highbrow, providing a standardized framework to streamline the visualization and discovery of annotated resources, not only in the digital text space but for other media formats such as video.

For example Professor Bob Kegan, Meehan Professor of Adult Learning and Professional Development at GSE in his approved HILT project proposal "*The lecture in 21st century learning: Reconstructing and revaluing our oldest teaching asset*" would like to explore innovative ways of "flipping the classroom" by having students engage in active viewing and annotate or tag fragments of lecture video topics before class. Then he would like to look at an aggregated display of all the lecture video marginalia, thereby quickly determining the places of greatest misunderstanding, controversy, learning-richness, and potential amplification. Finally, he can then enter class to engage his students in the material of his lecture at a whole "second wave" level, already knowing the most valuable veins to mine.

Although several video annotation clients are currently available within Harvard's LMS for Professor Kegan to use in his "flipped classroom" (see iSites *Video Collaborative Annotation Tool*, *Video Publishing Tool*), none of these tools offer a way to visualize and analyze an aggregated display of all the lecture video marginalia. One solution would be to interface the CATC API between existing tools such as the *Video Collaborative Annotation Tool* and *Highbrow*, therefore offering an accessible code mashup option instead of a potentially costly customized tool development alternative.

A further example for interface potential of the CATC API could be as a shared interoperable hub with the *Connected Scholar* tool could expand the idea development and notes features for texts in *Connected Scholar* to include audio, image and video resources as well. We met with Kimberly Hall of the *Connected Scholar* team to discuss the potential for the CATC as an interface and she saw that there would be much to be gained from the CATC.

Required resources

Phase I: Alpha Prototype development
16 weeks @ 35hrs/wk x 1.5 FTE @ $38/hr = $31,920

Managers time over 16 weeks (not contemplated in estimate):
Paolo Ciccarese - 5 hours/week
Phil Desenne - 3 hours/week
Marty Schreiner - 0.5 hours/week (consultation/review)

Phase II: QA, field testing and prototype evaluation
12 weeks @ 35hrs/wk x 2 FTE @ $38/hr = $31,920

Managers time over 12 weeks (not contemplated in estimate):
Paolo Ciccarese - 3 hours/week
Phil Desenne - 2 hours/week
Marty Schreiner – 1hour/week (survey design/testing/reporting)

Total funding requested = $63,840

Benefit Impact and Outcome Measurement (Failure/Success)
The number of projects and tools built for teaching, learning and research that will connect to the CCKH APIs will contribute to the measurement of the project impact.

In addition, we are planning an on-the-field assessment of major use case within the Harvard community. Several candidates, recipients of Library Lab and HILT projects (outlined above) and teaching faculty have already expressed interest in implementing this API within their own tools and projects. Field surveys will be conducted in the second phase of the project to qualitatively determine if the implementation goals were satisfactory. We will also perform quantitative evaluations by analyzing the number of annotations produced and shared across disciplines.