Annotation as Process, Thing, and Knowledge: Multi-domain studies of structured data annotation

W. John MacMullen

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360 http://www.unc.edu/~macmw

Following Buckland's (1991) work on the nature of information, this paper characterizes the multi-faceted concept of 'annotation' as process, thing, and knowledge. This typology is then used to enumerate general research questions for the exploration of annotation in arbitrary domains. Our research team's investigation of annotation of structured data in specific domains and user groups is described, including library catalogers, musicians, historical geographers, web users, statistical analysts, and biomedical researchers.

Introduction

The term 'annotation' bears a variety of meanings depending upon its context of use. In popular or vernacular use, an annotation is frequently defined as a comment or explanatory note in a printed text. Students encounter this type of annotation in textbooks or other types of readings (often historical texts), where unfamiliar words or concepts are explained. Another common instance is of the annotated bibliography, where lists of references are provided with context, explanations, and relationships.

In specialized vocabularies, meanings can be similar to these or vary significantly. In historical and religious scholarship, annotations can provide contextual detail about primary sources, or describe interpretations or differing perspectives. In the legal and governmental domains, annotations often provide references to relevant instances of an abstract concept in practice, such as court decisions associated with particular statutes, but the term can also refer to quite lengthy documents about specific cases. In some medical and clinical journals (especially in psychology and psychiatry), an annotation is an article-length document (often commissioned) that provides a review or synthesis of research about a particular topic (see, e.g., Viding, 2004).

In molecular biology and genomics, annotations are closer to what is often called metadata: terms and phrases used to describe an underlying resource (such as raw biological sequence data) with regard to its structure, function, location, and provenance (e.g., Stein, 2001). The majority of annotation-oriented research in the biomedical domain is focused on the problem of automatically deriving and assigning high-quality annotations to large databases of gene and protein sequences in order to understand single genes or organisms, and to aid in the recognition of cross-organism similarities of multiple molecules.

In information and library science (ILS), and in computer science, annotations are studied in terms of content- and process analysis (e.g., Marshall, 1998), as well as system design and functionality. Studies of content and process include instances of both authorial and reader-created annotations in textual and non-textual forms, including markings such as underlining and highlighting, as examples of sense-making and other motives and behaviors. Examples of system functionality include image and video annotation (e.g., Mu & Marchionini, 2003); annotation capabilities in collaborative systems (e.g., Mu, et al., 2003); question-answering IR systems (e.g., Prager, et al., 2000); and multiple types of manual and automatic annotation of text and audio sources in computational linguistics (e.g., Bird & Liberman, 2001). In the Internet community, annotation may mean anything from creating hyperlinks among distinct web pages, to assigning metadata to documents, to adding scholarly interpretations to existing hyperdocuments. There are more than 10,000 results for the term 'annotation' within the World Wide Web Consortium's (W3C) website alone.

A general typology for research

Given the range of extant conceptions of annotation as described in the preceding section, in order to study the creation, management and use of annotations in any arbitrary specialized domain, a multi-faceted definition of the concept is needed. If we assume that annotations are a form of information object, then following Buckland (1991), we can operationalize the compound concept 'annotation' into a typology of annotation-as-process, annotation-as-thing, and annotation-as-knowledge.

Annotation-as-process (A_p). As a verb, annotation is a process that has the function of creating or modifying an information object called an annotation. The study of annotation-as-process is the study of the ad hoc or recurring actions by which annotations are created, maintained, and used, by both human and non-human actors. These activities range in scope from individual personal annotation behaviors to automated annotation techniques to organizational workflows and practices that influence the process of annotation.

Annotation-as-thing (A_t). As a noun, an annotation is an intentional and topical value-adding note linked to an extant information object. 'Intentional' constrains the definition to purposeful notes and excludes artifacts such as accidental markings. 'Topical' limits the definition to only those annotations relevant to the underlying information object or use context, excluding such artifacts as graffiti or marginalia that are unrelated to the annotated item. 'Value-adding' implies that the presence of the annotation provides something of worth that is not present in the underlying object, such as an explanation or a reference. The term 'note' is purposely ambiguous, as an annotation can take many forms, including handwritten comments or sketches on printed pages, cells in spreadsheets, or fields in databases. 'Links' may be manifested in many ways, ranging from direct physical insertion (underlining on a printed page), to physical attachment (a note attached by paperclip to a page), to a hyperlink between two objects that reside in different information systems. Since an annotation is only an annotation in relation to some information object, the underlying object must exist prior to (or come into existence at the same time as) the creation of the annotation. 'Information object' is used as a broad term that encompasses such artifacts as documents (both printed and electronic) and database entries.

The study of annotation-as-thing is the study of the differing physical instantiations of annotations, and their properties and attributes, both alone and in relation to annotation-as-process and annotation-as-knowledge. The semantic meaning of the annotation is not considered in the study of A_t. No constraints are applied to the number of annotations an information object may have, and *n*-order annotations may be made to original annotations. (In other words, annotations themselves may have annotations, and so on.) The study of A_t is also concerned with the ability of an annotation to function as another type of information object in different use contexts, and with interoperability across contexts. What is an annotation in one use context, for instance, may be operationalized as a metadata element or an index term in another.

Annotation-as-knowledge (A_k). A_k is the intellectual component of an annotation, distinct from its physical manifestation (A_t). Knowledge is embedded in annotations, as it is in other information objects. The study of annotation-as-knowledge focuses on semantic meanings of annotations rather than their physicality. A_k is the 'why', not the 'how' of A_p or the 'what' of A_t.

The study of these three facets of annotation-as-concept helps to inform our basic understanding of information seeking and use behaviors associated with annotations, as well as the development of intellectual and physical tools and systems for the creation, management and use of annotations within their appropriate contexts. Figure 1 illustrates a generalized model of the components of annotation-as-concept.

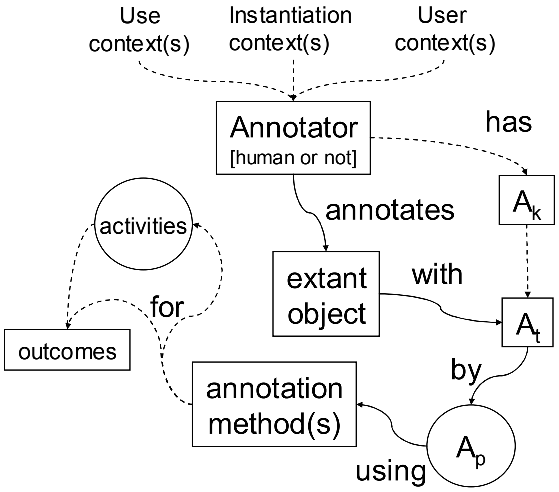

Figure 1. Components of annotation-as-concept

Research Questions

In the projects described below, the preceding facets are currently being explored in a variety of domains. A common set of research questions is being employed, which vary according to the situational context. The questions follow the typology in the previous section, and cut across dimensions such as those defined in Marshall (1998). In general:

- Process (A_p) questions explore the purpose and value of annotations: Why is annotation performed? What value or utility does the annotator create or derive from the annotation activity? Is it an end in itself or an intermediate step towards another goal or ongoing process?

What are the steps and the workflow in the process? What personal knowledge, skills and abilities and organizational assets are involved? Is training required to create annotations? Can someone unfamiliar with the system understand the annotation? Is there short-hand or coding involved? Is the annotation voluntary or mandated? Are annotations reviewed for accuracy, timeliness, completeness, or other quality facets? If so, what characteristics and attributes are reviewed? Using what criteria? Is the process generalizable to others performed by this researcher or other researchers? What is done with the annotations? How are they used and by whom or what? What processes take annotations as input or produce annotations as output? When in the life-cycle of the underlying object are annotations created? Can and should the processes be improved, and if so, how?

- Object (A_t) questions explore the structure and function of annotations as artifacts: What are the properties and attributes of the annotation object (e.g., provenance, format, permanence, relationships or linkages to other annotations)? Are standard formats or styles used? Can the annotations be characterized into types? Is a controlled vocabulary or domain-specific ontology employed for terms used in annotations? Would storage in a different format or medium allow higher levels of functionality or different kinds of utility? Is an annotation viewed as another type of information object in other contexts or under other conditions (e.g., as metadata)? How is the relationship between object and annotation instantiated? Is the annotation stored separately from the underlying object? Is the content instantiated in a form in which automated resource discovery, inferencing, or other types of processing could be employed?

- Knowledge (A_k) questions explore the meaning of the annotations and their intellectual relationships to other knowledge: What is the level of specificity of this knowledge? Is this knowledge related intellectually to that within other areas of this work? Does this knowledge have utility for other activities? Where can users go to find related knowledge, including that of broader or narrower specificity? How do people make sense of annotations?

The selection and formulation of these questions depends, on the situation under study. To address this variation, we are also investigating three types of contextual factors, as seen at the top of Figure 1: use context, instantiation context, and user context.

- *Use context* questions explore domain-level attributes and differences, such as academic specializations, and commercial industry segments.

- *Instantiation context* questions explore group-level attributes, such as impacts of group size on process and workflow.

- *User context* questions explore the practices of individual producers and users of annotations, such as: In what roles, job functions, or ranks do people create and use annotations? Are there demographic or sociological characteristics of annotators and users? Are there differences in skill-set characteristics between annotators and users?

Methods

In using the preceding research questions to investigate the domains below, a variety of research methods are being employed, including structured and semi-structured interviews, surveys, task analysis, the critical incident technique, and content analysis.

Current Work

Cataloging

Librarians who catalog resources frequently make annotations on and about the underlying works. The objective of this work is to characterize the decision-making processes and challenges that emerge when mapping either paper or online information resources to a structured vocabulary. The analysis will also explore the role that collaboration plays during cataloging process.

Luo et al. (2005) conducted semi-structured interviews with catalogers, and a content analysis of more than 2,700 annotated catalog records of an online consumer health resource. The analysis of the records revealed that establishing the geographic scope of the online resources, and pairing the subject headings to each service provided are particularly challenging.

The findings from these analyses will be used to develop a set of functional specifications that would enable a cataloger to overcome the challenges typically encountered during the cataloging process. In addition, the findings will inform the development and evaluation of an automated cataloging system.

Blake et al. (2005) conducted a qualitative content analysis to characterize the communication patterns that occur between catalogers as they assign controlled terms in the same consumer health information resource. Facets explored focused on annotation content, format, function, and changes in annotations over time. Results showed that catalogers most often discussed the topic, navigational scope, and geographical scope of an information resource. Annotations were most often in the format of a statement rather than a question or an answer. Catalogers made annotations as reminders to themselves or other catalogers, to reach consensus, to log an action, or to issue a request.

Music

Musicians, conductors, and composers annotate musical scores for a number of reasons: musicians annotate as a means to enhance memory and achieve reliable, consistent performance. Conductors annotate for the purpose of learning the score, and making decisions regarding supervision of the orchestra playing the work. Composers annotate scores of other composers in order to conceptualize and internalize ideas, themes, and methods for their own work. This project is examining annotations on musical scores of these three groups, at three different skill levels: amateur, college level, and professional; choosing musician groups that include conductors (orchestras) and those that do not (string quartets, for example).

The intention of this study is to investigate whether patterns and relationships exist between different user groups' annotation behaviors and needs; and to see if users at different skill levels annotate differently. An additional research focus is to learn whether any sections or musical attributes are commonly annotated across skill level and user group.

Although still in its early stages, preliminary findings suggest that annotation does evidence performance-related interaction between and among musicians, and is an effective means to identify those musical characteristics that are both important and variable across musician type and skill level. Finally, initial interviews with musicians and conductors provide some insight into the nature and importance of musical annotation. In addition to 'community of practice' and knowledge representation issues, the findings resulting from this study will argue for the importance of more robust and user-defined annotation facilities in the development of musical digital libraries and archives (Winget, 2005).

Historical geographers

Historians who work with maps and other structured geographic data frequently make annotations about primary sources, such as land surveys and other documents related to the ownership and conveyance of property. Ruvane & Dobbs (2005) investigated an historical geographer's use of annotations to create a multi-media time-based map illustrating land occupation in the North Carolina Piedmont region using a geographic information system (GIS). The geographer's objective was to demonstrate the influence a prominent transportation route, the Indian Trading Path, had on settlement patterns during the later half of the 18th century and the consequent emergence of today's urban centers. The project explored annotation facets such as different descriptive entity types that are captured as evidence for or against certain geographic boundaries. In subsequent work, Ruvane (2005) explored in more detail the multidimensionality of annotation within the context of historical geography, building on work by Marshall (1998), and exploring related concepts of information seeking in context.

Web usability for annotation

As described in the introduction, annotation creation and use in online environments is growing in scope and complexity. Fu et al. (2005) investigated the needs Web users have to make annotations for their personal use when they view Web pages. Three forms of annotations observed on printed documents – text selection and emphasis, link building, and document re-segmentation – were examined in the Web environment. An exploratory study shows that text selection and association building through notes or symbols remain the dominant forms of annotation on the Web, while structural annotation (re-segmentation) and layout annotation (change of font, color, etc.) are also prevalent. The study also investigated users' preferences for the tools designed to facilitate Web annotation practices. Findings suggest that usability is of utmost importance when developing Web annotation tools, and that under the current technical conditions, users welcome lightweight annotation functions built into standard Web browsers.

Social networking

Ciszek & Fu (2005a and 2005b) explore social hyperlinking in weblog or 'blog' environments as a form of annotation. A small group of regular bloggers were interviewed to determine bloggers' individual motivations for creating and maintaining blogs, and to assess their motivations for the creation of specific individual hyperlinks in their blog entries. This information was combined with demographic and geographic information for analysis and for the creation of a typology of author motivations for hyperlinking.

Government statistics

The U.S. Federal government collects, analyzes, summarizes, and publishes a large volume and variety of statistics. This work is performed by multiple agencies, each a large organization, using highly structured and formalized processes with very specific outputs. Quality control and documentation are of critical importance. Data are collected from numerous sources (individuals, households, businesses, institutions) using a variety of surveys and techniques, most of which have some facility for additional notes to clarify responses. These data are aggregated and processed centrally by a distinct set of people and systems that also include the possibility of new annotations. During the overall workflow, the data may be recoded, merged, split, re-analyzed, and presented in different end products, which are used by a large number of governmental, business, and news users. In this complex environment, there are multiple roles for annotators and users; multiple people within multiple agencies have cause and ability to make, change, and use annotations, and different formal and informal practices have evolved. Interviews with several statistical agency personnel have begun to reveal the roles that annotation plays in the overall flow of statistical information in government settings.

Biomedical research

Scientists and researchers in the biomedical domain use a variety of structured data sets in their work practices, ranging from simple spreadsheets and tables to extremely large databases with millions of records. In biomedicine, annotation as a process can range from informal, ad hoc notations by and for individuals, to formalized workflows as part of a larger-scale 'curation' process for wider audiences where value is added to raw data through both physical and intellectual linkages. This project investigates both the annotation behavior of researchers and the characteristics of their underlying annotations. Variation in annotation creation and use is explored through user contexts such as research role, job role, and functional role. Variation in annotations as artifacts is explored through content analysis.

A pilot study (MacMullen, 2005) examined annotations in nine model organism databases and the Gene Ontology (GO) to assess the quantity and types of explicit and implicit linkages between organisms. Despite having varying database implementations and interfaces, the model organism databases had similar annotation processes, content, and knowledge. While all databases had the potential to be linked to all others via GO, only some databases had non-GO links to others. This may be due in part to a lack of biologically significant relationships among some of the organisms.

Summary

The operationalization of the concept of annotation into a typology of process, thing, and knowledge, in conjunction with contextual information, has provided a framework by which annotation can be investigated within arbitrary domains, with the ability to compare outcomes across those domains. To date we have shown that formal and informal mechanisms for annotation exist in multiple contexts, and that by adding communication facilities to the standard tools promotes collaboration as a formal mechanism. We have also experienced these in our own work on this project through our use of a wiki as an environment for collaborative research and project management.

ACKNOWLEDGMENTS

This work was partially funded by an unrestricted research gift from Microsoft Research to the Annotation of Structured Data research team in the School of Information and Library Science at the University of North Carolina at Chapel Hill, whose members contributed to this work: Gary Marchionini, Paul Solomon, and Catherine Blake, co-PIs; with team members Tom Ciszek, Xin Fu, Lili Luo, Cathy Marshall, Mary Ruvane, David West, and Megan Winget. The project's website is available at: `http://ils.unc.edu/annotation`.

REFERENCES

Bird, S. & Liberman, M. (2001). A formal framework for linguistic annotation. Speech Communication 33(1,2), 23-60.

Blake, C., West, D., Luo, L., & Marchoinini, G. (2005) Cataloging On-Line Health Information: A Content Analysis of the NC Health Info Portal. Paper submitted to AMIA Annual Symposium, under review.

Buckland, M. K. (1991). Information as Thing. Journal of the American Society for Information Science, 42(5), 351-360.

Ciszek, T. & Fu, X. (2005a). An Annotation Paradigm: The Social Hyperlink. To appear in the Proceedings of the ASIS&T 2005 Annual Meeting.

Ciszek, T., & Fu, X. (2005b). Hyperlinking: From the Internet to the Blogosphere. The 6th International and Interdisciplinary Conference of the Association of Internet Researchers, Chicago, IL (October 5 - 9, 2005) (in press).

Fu, X., Ciszek, T., Marchionini, G. & Solomon, P. (2005). Annotating the Web: An Exploratory Study of Web Users' Needs for Personal Annotation Tools. To appear in the Proceedings of the ASIS&T 2005 Annual Meeting.

Luo, L., West, D., Marchoinini, G.& Blake, C. (2005). A Study of Annotations for a Consumer Health Portal. Poster to be presented in Joint Conference on Digital Libraries (JCDL), Denver, Colorado (Jun 7-11, 2005).

MacMullen, W. J. (2005). Inter-database annotation linkages in model organism databases. In Proceedings of the 68th Annual Meeting of the American Society for Information Science & Technology (ASIS&T), Vol. 42, Charlotte, NC (October 28-November 2, 2005), (in press).

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In Proceedings of ACM Hypertext '98, Pittsburgh, PA (June 20-24, 1998) pp. 40-49.

Mu, X., & Marchionini, G. (2003). Enriched video semantic metadata: authorization, integration and presentation. In Proceedings of the 66th Annual Meeting of the American Society for Information Science & Technology (ASIS&T), Vol. 40, Long Beach, CA (October 19-22, 2003), pp. 316-322.

Mu, X., Marchionini, G., & Pattee, A., (2003). The Interactive Shared Educational Environment: User interface, system architecture and field study. In Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries (JCDL), Houston, TX (May 27-31, 2003), pp. 291-300.

Prager, J., Brown, E., Coden, A., & Radev, D. (2000). Question-answering by predictive annotation. In E. Yannakoudakis, N. Belkin, M-K. Leong, & P. Ingwersen (Eds.) Proceedings of the 23rd ACM SIGIR conference, pp. 184 – 191.

Ruvane, M. B. (2005). Annotation as Process: A Vital Information Seeking Activity in Historical Geographic Research. To appear in the Proceedings of the ASIS&T 2005 Annual Meeting.

Ruvane, M. B. & Dobbs, G. R. (2005). Interdisciplinary collaboration and database modeling for historical GIS: structured annotation for land grant research. Association of American Geographers, 101st Annual Meeting, Denver, CO (April 5-9, 2005).

Stein, L. (2001). Genome annotation: From sequence to biology. Nature Reviews Genetics, 2(7), 493-503.

Viding, E. (2004). Annotation: Understanding the development of psychopathy. J Child Psychology & Psychiatry, 45(8), 1329-1337.

Winget, Megan. (2005). Digital Preservation of New Media Art Through Exploration of Established Symbolic Representation Systems. Paper to be presented at the JCDL Doctoral Consortium, Denver, CO (June 7 - 10, 2005) (in press).