Tags and Dependencies: an Integrated View of Document Annotation

Clemens Beckstein
Institut für Informatik
University of Jena
D-07743 Jena, Germany
beckstein@minet.uni-jena.de

Harald Sack
Institut für Informatik
University of Jena
D-07743 Jena, Germany
sack@minet.uni-jena.de

Heiko Peter
Institut für Informatik
University of Jena
D-07743 Jena, Germany
hpeter@minet.uni-jena.de

ABSTRACT

Metadata provided by annotations are an essential prerequisite for the realization of the semantic web. The manifold of metadata uses on the other hand also implies an abundance of different annotation types and formats, each requiring a different semantic treatment of the data. For semantically rich documents this results in a hybrid mixture of metadata about one and the same document. Further metadata diversity arises, if more than one author contributes to the annotation as it is typical for the now popular social tagging systems. Documents however, despite this heterogeneity show common structural properties that can be classified as logical, conceptual, and referential. There are intrinsic dependencies within and across those structures that have to be made explicit. We argue that the dependency structure as an explicit annotation is essential for any semantically rich document in order to be better understandable not only for man but also for machine.

1. INTRODUCTION

The Semantic Web holds promises for an intelligent organization of and a selective access to a huge, world wide distributed store of information by providing standard means of formulating and distributing metadata in a way accessible to machines. Metadata are structured, encoded data that describe characteristics of information-bearing entities, as e.g., documents [2]. Metadata can be connected to documents by annotations. We distinguish between annotations of the document's author, as e.g., annotations that define the logical document structure, and annotations given by the document's user, as e.g., referential annotations or descriptive keywords provided by collaborative tagging systems. Users of documents range from casual readers to software agents that process documents on their way through the network. The success of the Semantic Web relies on the proper annotation of its information resources. For this reason, a lot of effort is spent on tools for the annotation of existing resources as well as resource authoring tools that provide annotation facilities.

Many information-bearing resources comprise complex annotation that can only be fully exploited if it is related to annotation from associated external documents. We claim that in order to cope with the corresponding heterogeneous annotation structure an integrated view of document annotation is mandatory. Despite of this heterogeneity documents show common structural properties. They can be classified according to their logical, conceptual, and referential characteristics. There are intrinsic dependencies within each of the three structures and also across them. These dependencies have to be made explicit in order to facilitate cross annotation reasoning that is necessary in order to make full use of the information stored in the documents.

The paper is structured as follows: Section 2 describes our integrated view of document annotation and shows possible applications. Section 3 then gives reference to related work and Section 4 concludes the paper with an outlook on future work.

2. TYPES OF ANNOTATION AND THEIR DEPENDENCIES

2.1 Documents, Tags and Annotations

From an abstract point of view a document is just a plain, totally ordered string of individually addressable document tokens. These tokens form the smallest units of the document that can be addressed by its very position in the document. The document string typically is interspersed with so called tags that are added to the document by its author and the other users. Tags come in different types corresponding to different uses of the tagged document parts. Their primary function is to associate distinguished parts of the document with metadata that are specific for these parts and can be put to use by processes that operate on them. Tags also carry information about their creators and thus enable personalized views on the document. In the following we will denote both individual tags as well as groups of tags that semantically belong together as annotations.

The most prominent example of a document is the text document, e.g. a textbook. The tokens of text documents are words and other document units like figures that are considered to be atomic. Structural tags are used to form higher level document units, as e.g., sentences from words, paragraphs from sentences, or chapters from sections. Another example are video documents. Video data can be encoded according to the MPEG standard [6]. There, the single pixels of a video frame as being the smallest addressable units can be considered as tokens. Pixels can be arranged within blocks that again are subsumed into macro blocks. Additionally, the MPEG-4 standard allows the definition of distinguished objects that can be arranged within a scene.

Recently, collaborative tagging systems (CTS) have become increasingly popular for annotating any kind of ressources. In CTS the user assigns tags to specific resources with the purpose of identification and reference. If the tags being assigned by other users are considered as well, resources might be discovered serendipitously by so called "tag browsing" [7]. CTS enable additional annotations that in difference to traditional authoring systems are provided by the users and not by the author.

Documents can be structured along three dimensions: they exhibit a logical, a conceptual, and a referential structure. These structures are induced by the linear flow of the smallest document units, by the semantics underlying the document tags, and by the dependencies that exist between these tags. They are therefore determined not only by the author but by all the programs or persons that process the document and add their own specific tags to it.

2.2 The Logical Structure

The logical structure of a document captures the part-of structure of the document units defined by the tags and reflects the total order of the smallest document units. From an abstract point of view it is just an ordered tree. The nodes of this tree represent document units. The link between a child node and its parent node reflects the fact that the child node represents a document unit which is a direct constituent of the parent's document unit. The order of the tree follows from the total order of the smallest document units: for any two nodes n_1 and n_2 on the same level of the tree, n_1 precedes n_2 whenever all the smallest document units belonging to n_1 are located before all the smallest document units belonging to n_2 in the document string. Different views held by different tag creators in general represent different logical structures of the document. But all the part-of relationships of these structures will be compatible with the logical structure of the total order of the token stream.

Structural tags need not always be specified explicitly, as e.g. by the delineation of a chapter in a book via a corresponding mark up. They may also be present just implicitly via the format or layout of the document—like the word boundary (via punctuation and white space) or the scope of a paragraph (via a separating line) in a text. Explicitly specified structural tags usually define not only a higher level document unit but also a name for this unit—e.g. a short, summarizing title for a book chapter—that is meaningful to the person that added the tag to the document.

The part-of tree along with explicit or implicit names for the document units can be used for document navigation since it gives rise to complex absolute and relative addresses. The absolute address of a document unit is given by the shortest path from the root of the part-of tree to this document unit. The relative address of a document unit u_2 wrt. to a higher level document unit u_1, which it is part of, is given by the path leading down from u_1 to u_2. The ability to systematically form complex addresses of document units is essential for the referential structure of a document.

An interesting complex annotation can be derived from the logical structure and the explicit structural tags of a document by traversing the corresponding ordered part-of tree starting at the root, depth-first and left-to-right, until a given level of document granularity is reached. Collecting the names and addresses of the document nodes belonging to the visited nodes results in the familiar hierarchical organized list known as table of contents (TOC). For text documents the list elements of course correspond to headings of document units and the addresses to page numbers starting the respective units. In analogy to the TOC for video documents an annotated list of consecutive video segments (scenes or chapters) can be considered. Video segments can be identified automatically within the MPEG-4 encoding of the video data, while annotations describing the video segment can be extracted from MPEG-7 metadata [1]. In the same way as the author of a text document identifies a section by putting a numbered heading on top of that section, the author of a video document can mark up a video segment by adding a MPEG-7 scene description.

2.3 The Conceptual Structure

Documents also exhibit a conceptual structure which can be considered as its ontological skeleton [5]. The conceptual structure captures all the concepts that are subject of the document along with the predominant relationships holding between them. One important example for a relationship between two concepts is concept subsumption. This relation allows to automatically generate terminologies ranging from simple flat ones like glossaries to complex ones like sophisticated concept hierarchies.

The author and other users of a document usually specify only a small fragment of the conceptual structure of a document via tags. Most of the conceptual structure is contained only implicitly in the document and requires natural comprehension in order to be made explicit. For this explication key units of the document have to be identified and associated with the (names of) concepts applying to them. The resulting concepts then have to be related to each other and to concepts of other documents by further annotations until a rich enough fragment of the document's conceptual structure is uncovered. For this process again the referential structure of the document is an important resource.

Tags that contribute to the conceptual structure of a document can range from very simple, as e.g. index entries, to highly complex structured ones. Index entries explicitly specify both the association of the document unit with a concept as well as the exact position of the involved concept names. An indexing process can then combine the tags from the logical and the explicated conceptual structure of the document to derive another complex document annotation—the familiar hierarchical list of concepts along with their occurrences, which is called the document's index. Computational tools that take into account general knowledge about indexing and document related ontologies can considerably improve the overall quality of a document's index [10].

In analogy to index entries, for video data MPEG-7 descriptions can be used to annotate video segments with conceptual information. In a limited way even an automated annotation with conceptual information is already possible: Automated annotation tools are able to identify cut points between consecutive scenes and to supply visual descriptions on different levels of abstraction [3]. The transcription of the audio content that is part of the video data can also serve as a basis of conceptual content annotation [11].

Conceptual annotation of course can be much more complex than the provision of a list of index entries, where both the identification of a concept and the relationship between two or three concepts is determined by a single tag. It usually requires both tags that just associate document units with concept (names) and a formal language that allows to express the relationships between the concepts named this way. Languages suitable for this annotation task can be very complex and even be undecidable as the family of Web Ontology Languages shows [8].

2.4 The Referential Structure

The third important structure that a document exhibits is its referential structure. The cross reference relation of a collection of documents is defined as the set of all pairs (u_1, u_2) of documents units, where document unit u_1 mentions document unit u_2. If u_1 and u_2 belong to the same document, then the pair (u_1, u_2) is called an internal link from u_1 to u_2; otherwise it is called an external link.

The cross reference relation of a collection of documents can be visualized as a directed graph, where the nodes represent document units and the directed arcs the links existing between them. Whenever there is a directed path in this graph leading from a document unit u_1 to another document unit u_2 then this suggests that the concept being expressed in u_1 is in some sense dependent on the concept u_2 is about. What kind of dependency is meant exactly depends on the types and contents of the documents involved.

Usually only a small part of the referential structure of a document collection is specified via tags. As for the conceptual structure, most of it is contained just implicitly in the involved document units and would require a deeper understanding of the respective document parts in order to be made explicit. An author or other user of the document collection can add a link (u_1, u_2) to a document unit u_1 by placing a tag somewhere in u_1 that contains an address referring to u_2, which is formed in accordance with the logical structure of the document containing u_2.

Internal references of a text document refer to document units like footnotes, other chapters, or figures. Familiar examples of external references of text documents are links to other documents or part of them. It is also typical for complex text documents that the referential tags belonging to different sorts of cross linked document units are compiled into complex referential annotations like the table of figures, the list of endnotes, the list of notes, or the list of external links known under the familiar name "references".

2.5 The Structures in Concert

The logical, the conceptual, and the referential structure taken together open up new kinds of annotation based document applications that would not be possible by just using one of them in isolation.

A prime example of such an application is a computer generated reading tour through a collection of documents, which reflects the metadata from the authors and the tagging information provided by the community of the users of these documents. Simultaneously, such a tour, e.g., could be used by an e-learning system to automatically generate a sequence of learning units that have been identified as relevant for a course and that were previously semantically annotated for this purpose by a human teacher. Another, simpler use of a tour, for the users of a single but complex document, would be the suggested reading (dependency) graph as it is found in the introduction of certain voluminous textbooks. For a sound tour document units that—according to their referential structure—are dependent on each other should be visited only after the ones they depend on. The suggested reading tour should also guarantee that concepts document units in later parts of the tour are about should have been defined earlier in the tour. This can be achieved with the help of the conceptual structures of the collection. Finally, the default order on document units at the same level of the logical structures of candidate documents for the tour can be used to constrain the reading order, where no conceptual or referential dependency determines their precedence.

The documents' logical and conceptual structures together also provide a basis for a goal-oriented selection of document units for the reading tour. The conceptual structures can be used to collect the relevant document units, i.e. those that are mandatory to cover all the concepts the user is interested in. The logical structures of the candidate documents then can be used to keep the size of the relevant documents units as small and concise as the part-of relation of the involved documents permits. User provided annotations also allow the generation of personal tours tailored to the information needs of different users.

Another application dependent on an integrated view of document annotation is collaborative authoring. For this application the logical and the conceptual structures of the involved documents have to be merged. Obviously, considering only logical and conceptual dependencies within the single documents will not be sufficient—cross document dependencies have to be taken into account.

3. RELATED WORK

We are not the first to recognize that the different metadata, i.e. the different sources of knowledge about a semantically rich document or group of documents that are provided by the authors and users each exhibit intrinsic structure and depend on each other. They form what is called a hybrid representation in the field of knowledge representation [9]. The key problem with hybrid knowledge representation formalisms is to guarantee that reasoning processes operating on hybrid knowledge structures do not only reasonable work wrt. the individual knowledge types but also across the borders of different knowledge types.

Topic Maps [4], e.g., are a knowledge representation formalism that attempts to express (aspects) of hybrid structures of this kind. They combine a semantic net like conceptual representation with a reference structure for the resources that this semantic net is about. Topic Maps could therefore provide a basis for the representation of the conceptual and referential structure of a document. They miss however dedicated means for the representation of the logical structure of a document. In addition they do not even guarantee that processes operating on the conceptual or referential structure alone behave according to their underlying semantics—two deficiencies they share with early semantic nets [14].

Topic Maps also do not answer the question of how to design annotation that is suitable for a given collection of documents. But there are a number of other tools, which support this by uncovering parts of a document's structure. Well known representatives are automatic document annotation systems that attempt to extract the conceptual structure of a document using information retrieval techniques. E.g., in [12] the video recording of a lecture is synchronized with the lecturer's presentation to extract semantic annotation that is utilized for a content based search within the video data.

In [10] we describe an approach that combines general knowledge about indexing, document related ontologies, and structural knowledge about a given single text document for the computer supported generation of a high quality document index. The corresponding SmartIndexer system internally uses a directed graph—the so called Index Graph—that shares structural resemblances with Topic Maps. The Index Graph is layered into two subgraphs: a structure graph and a document graph. The structure graph represents certain aspects of the conceptual and referential structure and the document graph the logical structure for the document to be indexed. Another approach that uses a part of the conceptual structure of a document for indexing purposes is documented in the work of Shabajee et al. [13]. They propose a system that supports the automated annotation of multimedia database indexes utilising extensible controlled vocabularies. Their approach distinguishes users by their level of expertise and integrates an according rights management.

4. CONCLUSION

We have shown that document collections despite their semantical heterogeneity possess intrinsic logical, referential, and conceptual characteristics and that complex dependencies exist within and across the document structures carrying these characteristics. We have also argued that these characteristics as well as their interdependencies have to be made explicit and should be maintained along with the documents that carry the underlying metadata. The corresponding dependency structure can not only be rather useful for text documents, as we have sketched in this paper, but also for other types of documents.

Adequately processing heterogeneous document collections based on their distributed and hybrid metadata—by the very nature of this endeavour—requires the explicit maintenance of the annotation structures of the involved documents along with the dependencies that exist between them. Difficult as this is, it does promise to open up the door to new and exciting applications that seem to be impossible without.

5. REFERENCES

[1] S. F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 Standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695, 2001.

[2] W. R. Durrell. *Data Administration: A Practical Guide to Data Administration.* McGraw-Hill, 1985.

[3] S. B. et al. Semantic annotation of images and videos for multimedia analysis. In *ESWC*, pages 592–607, 2005.

[4] L. M. Garshol and G. Moore. ISO 13250-2: Topic Maps — Data Model. Final draft, ISO/IEC, 2005.

[5] T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisitation*, 5(2):199–220, 1993.

[6] ISO/IEC. Overview of the MPEG-4 Standard. *ISO/IEC JTC1/SC29/WG11 N2323*, 1998.

[7] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.

[8] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language: Overview, W3C Recommendation, 10 February 2004.

[9] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems.* Number 422 in Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, 1990.

[10] H. Peter, H. Sack, and C. Beckstein. Document Indexing – Providing a Basis for Semantic Document Annotation. In *XML Tage*, Berlin, Germany, 2006.

[11] S. Repp and C. Meinel. Semantic Indexing for Recorded Educational Lecture Videos. In *4th Annual IEEE Int. Conf. on Pervasive Computing and Communications Workshops (PERCOMW'06)*, 2006.

[12] H. Sack and J. Waitelonis. Automated Annotations of Synchronized Multimedia Presentations. In *In Proc. of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, CEUR Workshop Proc.*, 2006.

[13] P. Shabajee, L. Miller, and A. Dingley. Adding Value to Large Multimedia Collections Through Annotation Technologies and Tools: Serving Communities of Interest. In *Museums and the Web 2002: Selected Papers from an Int. Conf.*, Archives & Museums Informatics, Boston, USA, 2002.

[14] W. A. Woods. What's in a Link: Foundations for Semantic Networks. In D. Bobrow and A. Collins, editors, *Representation and Understanding.* Academic Press, 1975.