

Annotating as a document management tool

O. Mazhoud, E. Pascual and J. Virbel

118, Route de Narbonne, F31062 Toulouse, France

KEYWORDS: digital library, dynamic annotating, reading activities

AFFILIATION: Institut de Recherche en Informatique de Toulouse (IRIT)

E-MAIL: mazhoud@irit.fr
pascual@irit.fr
virbel@irit.fr

FAX NUMBER: 61 55 83 25

PHONE NUMBER: 61 55 67 64

I Presentation

Current technology (scanners, OCR) is paving the way for the creation of vast machine-readable document holdings (digital libraries, technical full-text databases, CD-ROM, etc.). At the same time, the range of potential users is also widening considerably. It appears that under these conditions, very general language engineering tools may be both too powerful and deficient in providing for all conceivable user situations. In addition, personalized, reader-oriented document management tools such as annotation have to be designed for running reading products.

To reach this aim, three partners (BNF [Bibliothèque Nationale de France], AIS [Advanced-Information Systems] and IRIT) have already collaborated in the past with regard to the definition of a strategy for creating the machine-readable holding for the BNF and specifications for the CARS (Computer Aided Reading Station) [Chahuneau & Al, 1992]. This work is now developed as a part of the European MEMORIA project (Multimedia Electronic MemORIES At hand) and is used as a background for this article.

II Electronic annotating

Nowadays, researchers pay a particular attention to electronic annotating [Nielsen 86], [Stiegler 94], [Virbel 93]. This is due to the convergence of two principal factors:

(1) from a technical point of view, and particularly in a context of hypertexts, users are allowed to widen their documents by adding different individualized augmentations such as characterizations of passages, attachments of various comments, links between passages, etc.

Since all annotations could be memorized by a system, they present the two following major advantages, compared to traditional form annotating

(with paper and pencil):

- The set of annotations itself constitutes a material to be managed and exploited subsequently, as long as the tools necessary for exploitation are available.
- Re-reading and later consultations of the document can actually be performed through annotations.

(2) at the same time, the technical feasibility of creating very wide corpora of digitized documents (from paper books and catalogues) and the profile broadening of potential users of these corpora, favour personalized management modes of these corpora. Therefore, annotation seems to provide a mode of personalization of the document through its reading, since it is qualified as a highly individualized activity.

In this context, we characterize this activity as a *Dynamic Annotating* contrary to the static character qualifying it in paper context. The characterization as dynamic is due to the following main reasons:

- firstly, as in the classic context, the annotating activity is contemporary with the reading process and more generally all reading concerns;
- secondly, various types (discursive, graphic, sound, etc.) of annotating trail themselves constitute a material to be managed and exploited as far as suitable tools are available;
- thirdly, the re-reading of the text could be organized from and through annotation. So annotation could be considered a new access mode to the text (such as the table of contents, the alphabetic index, keywords, etc.);
- fourthly, annotating can provide users with a sophisticated *experimental* tool for reading: for example searching for new passages sharing some properties with a given one, or creating passages automatically if they have not been created, may be considered a relevant element of the dynamic character of annotating;
- finally, if annotations are well managed, and the expansion of the system is well controlled, then this evolutive system could be a considerable aid for researchers who study the comprehension phenomenon associated with the cognitive activity of reading.

III Elementary typology of annotation and the composition problem

Experiments carried out within the scope of the BNF, have proved the importance of different modes of annotation. It seems to be a form of writing which is intrinsically bound to reading, as

well as a method of management of the reading itself and therefore the “re-readings” or subsequent consultations of Digital Library.

According to these experiments, functions generated by the annotating activity seem to be numerous. They particularly concern memorization; capitalization of the reading’s results; indirect dialogue with the author of the text under consideration; possibility of communication between readers; management of the subsequent consultations and re-readings; scheduling of the operations to be carried out aside from purely reading. From the observations collected from professional or academic readers (who were in fact mainly researchers in the fields of Humanities and Social Sciences) using the sources of the BNF, we defined eight principal elementary events of annotation:

- (1) Organize into a hierarchy: Attach degrees of judgement according to importance, representativeness, etc.
- (2) Create an architecture: Add on by explanation of structural elements (e.g.: highlight items for a purely discursive listing).
- (3) Contextualize: Create a passage which is relevant for semantic apprehension of a term or expression.
- (4) Schedule: Plan operations to be carried out aside from purely reading (e.g.: “to re-read”, “to translate”, etc.).
- (5) Reformulate: Modify the content of the text.
- (6) Comment: Attach some text to a point, or a passage of text.
- (7) Document: Attach a document (ex: picture, sound, film).
- (8) Correlate: Establish footnotes and references between points, passages, or zones of text or of another text.

After selecting an object (called a passage) from a document, the reader assigns a type of annotation to the passage. In this context, annotated passages can be defined either by direct selection (for example by using the mouse), automatically from other passages which have already been created, or by using markers, if they are present.

Concerning the question of composition, we define a sophisticated model of complex annotating. In fact, we can say that the power of an annotating system can be measured by the intrinsic quality of the kind of annotations defined, and by their capacity to enter in composition in order to represent complex annotating acts ([Mazhoud et al 95]).

We define a complex annotative act as a combination of elementary acts with respect to an annotation composition mode. We distinguish the following modes:

- (1) “Accumulation”, which is tied to the spatial layout (inclusion, overlapping, etc.) of considered passages.
- (2) “Reapplication”, where a reader can create annotations on top of annotating acts already performed by the reader (e.g. to comment the act of criticizing a given passage rather than the textual content tied to the critique).
- (3) “Macro-annotating” is the means of composing more than one elementary event in the same annotating act.
- (4) “Chain”: as annotating events can be classified into two classes, characterizations and attachments, this mode concerns only the second class. Indeed, events that produce new texts can be the object of a continuity of annotation performing.

As some combinations of elementary acts are not relevant, a CARS must provide users with a composition control unit which ensures that the reader’s compositions are well formulated. To be sure, this control acts only on a “syntactical” level.

IV Annotating and reading activities

We present in this communication the *Annotating* activity and its relationships with other reading activities, in particular *Mark-up*, *Prospecting* and *Forming into corpora*. During the CAR, it appears that these activities are strongly correlated. An example of a reading session will be given to show this interdependence.

- (1) Mark-up i.e.: the association to the text of various markers denoting units of its logico-linguistic structure (sentences, paragraphs, acts, scenes, etc.). Here author-structure is in question, with regard to this sort of reader-structure where a part of annotation is composed.
- (2) Prospecting, which may mean the set of possibilities offered to carry out detailed investigations of the text, in lexical, syntactic, semantic and stylistic terms, etc.
- (3) Forming into Corpora, i.e. the classification of (segments or units of) texts, that is the composition of textual entities coming from various texts into new sets (corpora), arguments of textual operations.
- (4) Annotating: the reader defines relevant passages which will become arguments of textual operations.

It was seen above that annotating activity is defined as an event which consists of the selection of a passage in a document, and of assigning a type of annotation to this passage. Therefore, how can a passage be selected? Different methods, which are classics in reference theory, also run for referring to passages:

- direct reference:
 - showing (e.g. with mouse),
 - naming (proper name),
 - determining a set of properties;
- indirect reference:
 - determining a passage by properties shared with another passage.

As Mark-up and Forming into corpora functionalities may be represented and tailored in terms of SGML DTD's (e.g. TEI), it would be very interesting to also represent annotating activities and composition in the same frame. This hypothesis has not been evaluated for now.

V Towards an “experimental” reading

A reader-oriented document management system with the annotating functions described above must have the following main features:

- a form of memorization of different kinds of reading activities such as consultation, sorting, etc., with respect to reading sessions;
- at the same time, a form of “opportunism”, in fact every instantaneous idea or intuition must be immediately recorded, and its effects are systematically related to the set of the concerned area;
- a form of exhaustiveness, i.e. the possibility of formulating and testing in near real time various hypotheses that otherwise remain unchecked and, in consequence, the possibility of recording the results of these tests.

These features lead readers to a new way of reading. It is no longer a matter of simply following lines in screen, but rather of discovering new notions in the text by employing new concepts of approximations and controlled reiterations. This is what we call “experimental” reading.

References

- Chahuneau F., Lécluse Ch., Stiegler B., Virbel J. Prototyping the Ultimate Tool for Scholarly Qualitative Research on Texts. *Seme Conference Annuelle du New Oxford English Dictionary*, Waterloo, 18-20 Octobre 1992.
- Mazhoud O., Pascual E., Virbel J. Representation et gestion d'annotations. *Hypertextes et Hypermédias*, Hermes, 1995, 127-138.
- Nielsen J. Online Documentation and Reader Annotation. *International Conference on Work with Display Units*, Stockholm, 12-15 May 1986.
- Virbel J. Reading and Managing Texts on the “Bibliothèque Nationale de France” Station. In Delany P., Landau G. eds. *The Digital Word: Text based computing in the Human-*

ties, MIT Press, 1993, 31-52.

Stiegler B. *Machines à écrire et matières à penser. Genesis*, 1994.

Sperberg-McQueen G.M, Lou Burnard eds. *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, ACH-ACL-ALLC, 1990.