Two studies of user needs for bibliographic software

Peter Schulte-Stracke

13th May 2003

1 Introduction and Methods

In the following, I discuss two comparable studies of user needs [Knorr, 1998] and [Job-Sluder et al., 2003] which despite obvious differences had similar goals and results.

Knorr [1998] studied the impact of external sources on the academic writing process. Her dissertation precedes the recent OpenOffice study by about ten years, if one looks at the time of the empirical research. She did a survey of about 50 persons, mostly graduate students and post-graduates, some of whom were quite accomplished and experienced writers. Their works included sizable books and touched subjects like literature with considerable bibliographical requirements both from the size and from subject matter.

The survey in itself resulted in a lot of unremarkable statistics, but also some interesting observations. Her main contribution, however, lies in a couple of case studies where Knorr was able to observe her subjects while writing.

In their comparable but less extensive study [Job-Sluder, Hanek, and Zoang, 2003] restricted themselves to a smaller and more homogeneous sample. They proceeded by structured interviews.

We can state at the beginning the following main result of both studies: There is no strong pattern of bibliographical usage. Subjects used a variety of methods, often quite idiosyncratic ones and were both flexible and inflexible in a way that effectively precludes comprehensive 'solutions' for their needs.

2 Bibliographic description

During her work, Knorr did a survey of academic writing guides and found most of them wanting, in particular with regard to the use and preparation of bibliographical data.

It appears that the problems of bibliography (not only tool use) are generally underestimated. Some of Knorr's subjects had invested considerably in the database and sometimes programming, but generally the use of bibliographical tools is haphazard and amateurish in both studies.

Inadequate tools and understanding may easily lead the user into a blind alley. If a database has grown to a certain size, the user might feel that he cannot afford a switch to a better solution, in addition to any difficulty of breaking with deeply ingrained habits. Here it is important to understand that writing is for many people an important activity that (over time) has become part of one's personality and way of life. So that is a limiting factor and at the same time underlines the importance of sound foundations.

It is sobering to note that in both studies not only advanced features of bibliographical software were not used, but also the standard usage of adding references to a paper (BibTEXing), dating back to the sixties, was often enough rejected in favour of tedious manual methods, e.g., keeping the data in word processor files or generic 'databases' like *Filemaker*. From that it would appear that one important aspect of bibliographical software must be user education.

This stands in contrast to the idea that [Job-Sluder et al., 2003, p. 3] espouse, of raising the claim of the *Reference manager* to the heights of *Knowledge Management*. Although it is certainly possible and potentially interesting to consider the bibliographical description as a problem of knowledge management (cf. Fattahi and Parirokh [2002] and also [Svenonius, 2000]), it is an elusive aim, even in this restricted meaning.

3 Subject access

Only in exceptional cases is the user interested in the bibliographical description in itself. In most cases it is only a means for eventual access the item itself, and not for its formal attributes but for its content. The making available of instruments suitable for this purpose is known as *subject cataloguing*(or sometimes *indexing*, but v. *infra*) and its methods are:

Summarising helps with the overview by making descriptions shorter, more uniform and more focused. Abstracting is the best known special case, but as a general technique it is used everywhere. In the most radical sense, *genres* and *formats* reduce the whole of a resource to one word.

Indexing registers resources according to certain attributes; as in the preceding case, this is a very general intellectual technique, employed, e.g., if a book is filed under its author. Indexing is of very great importance, as it can be applied to:

- *keywords* as found in the title, an abstract, or elsewhere in the record or in the item itself
- *descriptors*, including subject headings, but more usually with great success normalised data extracted from the item; such as the microbiological species in the Index Medicus, or the statutes, in fine detail, cited in a decision or treatise
- finally cross references: indexing by cited or citing work, and many more. This is a field that has made great strides in the last time, allowing one to *navigate* through the bibliographical universe, but also serves as one means of automated indexing.

Classifying places the item into one or more classes, usually represented by a *notation*. Contrarily to what one might expect, such classifications are particular good at expressing the manifold aspects under which one might investigate the reality; thus it is not usually the *paradigmatic* or genus/species hierarchy that finds itself expressed in such a classification.

Subject headings express the contents of a resource verbally, usually employing terms from a *thesaurus* and joining them syntactically. The emphasis lies on grouping together works that share the main content in as small a class as possible – reserving a term like Mathematics for books about the whole of mathematics. Using and applying such a *subject heading language* is not simple (though it was originally conceived by Cutter exactly as a simple to use replacement for classifications), maintaining a thesaurus a major task. But a user may receive a lot of subject heading information from bibliographic databases, and perhaps want to extend its use to other resources.

Subject headings and classifications are neither fully substitutes of each other, nor are they complete opposites. They complement and profit from each other.

One of the unwelcome connotations of the term *subject cataloguing* is that it is the sole province of the library. That means, however, exaggerating the possible rôle of the latter, in fact wildly extrapolating from what is possible in descriptive cataloguing. For archivists it is on the other hand self-evident that whatever work they do in indexing etc. is only preparatory and assistive, and that it is the user himself whose task it is to find, order and interpret the sources. And so it is, in fact, universally.

As subject access is expensive and imprecise, it deserves some thought. One fundamental problem is its **perspectivity**: on the one hand, the division of labour in society and academia and more generally its structure have led to an unescapable pluralism (Luhmann), that must be respected by bibliography, on the other hand, users themselves often experience their grip on the resources as volatile, fading as soon as they begin another task. In addition it is a seemingly **paradoxical** task: in order to organise the materials one must already understand it, and for both students and researchers 'the owl of Minerva only flows in the dusk' (Hegel).

Thus the value of sophisticated methods of subject access is often questioned (since the middle of the 19th century at least [Svenonius, 2000]). Indeed, most participants in Knorr's study struggled with it, mostly relying on keyword search – and living with poor precision and recall. Sometimes, if the office space permitted, they classified by physically heaping papers on shelves and floor and putting their episodic memory to work instead of their computer. For others avoiding such a situation was exactly their foremost goal in using it.

Taking all this together one might wonder whether, e.g., for a law student, the use of a moderately developed classification, such as found in the table of contents of standard textbooks, together with indexing on statutes and cross references, would be a suitable, stable and not overly expensive way of organising resources, more suitable and in the long run less expensive than organising the stuff in binders, or on tables etc.

4 Work flow

Job-Sluder et al. use the unexpressive word *meta data* for both the subject cataloguing data, and also annotative and task-related data. But it is important to take a nearer look.

Knorr uses a more detailed schema:

1. The descriptive cataloguing data, as used for finding in libraries, citing (i.e. indirectly for finding), to a lesser degree for description per se.

   In her study, no use its made of machine-readable data, although she expected that its use would soon become widespread; it is thus surprising that the same applies to the recent study by Job-Sluder et al. as well. In the older study, someof the participants spent much time on collecting and correcting citations, reflecting perhaps their respective fields of study.

As little use was made in both studies of automatic reference formatting, little can be inferred about the requirements of this particular feature.

2. The subject-cataloguing data, for which Knorr coins the nice term »fachtextübergreifend«, and which she correctly sees as expressing something that is *not* simply contained in the item, but attributed to it by the user (in the last instance).

   This has been expanded upon in the section 3, so it may suffice here to emphasise the difference to annotations, as below.

3. The annotative data, either transcribed like abstract, quote, excerpt directly from the item, or formulated by the user. The purpose and point of view can serve to differentiate annotations from subject cataloguing data: they are 'subjective': more task-oriented, more reflecting the original purpose, less disciplined, potentially much longer.

   Here, of course, the question arises, whether or not this is the business of the reference manager. After all, these data can in itself become the object of the reference manager's services (take indexing of quotations as an example); it may slowly evolve into works of their own; and on the other hand whole books have been written as annotations.

   There is an interesting design problem here: on the one hand integrating for convenience and perhaps speed, on the other hand taking apart for simplicity, independent development, and extensibility. It is at the moment undecided which direction will prevail in the long run.

4. Task-related data, that is ephemeral by its nature; it is perhaps not so often stored in the database, although there are little problems with doing so.

5. Access providing data, in other words *holdings information*. Today this includes Internet addresses as well as call numbers etc. with their respective holding institutions, and – not to forget – corresponding information for items in the possession of the user: binders, shelves, but also file names on disk. This information, as minor as it may seem, is of great practical importance, so that it is somewhat surprising that in Knorr's study not a single person used the same recording scheme for internal and external items, many relying on their episodic memory for the former.

There are many ways to use the collected information during writing: much depends not only on the personality of the author, but also on the expected size of the work, the allotted time, and the degree of routine that is given. Nevertheless, it is generally possible to distinguish at least the following usage situations:

- Early: to build a reading list
- During planning and drafting: to build an argument
- During writing: to enter citations and quotations
- After writing: to check and correct citations (often neglected but mentioned by one participant of Knorr's study)

It is, of course, impossible to discuss all conceivable ways of using a reference manager application. It is, however, quite useful to analyse possible *interactions* between the various options a user has.

One obvious case is the interaction between the use of electronic storage for articles and the availability of more developed cataloguing for them; others, less obvious, exist between the degree of preparation for writing or the time horizon for a writing task, and the kind of queries that will be asked. In Knorr's study one could read between the lines that there are such contrasting ways of working, in the study by Job-Sluder et al. there is little detail given to help one in this respect.

By way of comparison the participants in Knorr's study had more diverse – and by far more conventional – ways of acquiring references for their work; however part of this may be due to the ten years that lay between both studies. They also seem to have spent more work on organising references, but again that may be an artifact of the presentation.

The post-production stage is not mentioned in the OpenOffice study, but by Knorr: this involved checking citations, but also maintaining cross references between one's own works and the database.

Notably absent from both studies are non-book materials, in particular archival sources. Perhaps they are in fact so seldomly used, perhaps they are so different as to be overlooked in this connection; equally possible that they only less visible because not easily processed by the more traditional tools. In the future it may be expected that the demand for integration will increase, though.

5 Miscellaneous points

1. It is important to distinguish between the following aspects, at least it will help keeping the discussion more focused:

- Data entry – *many* problems can be solved early on by paying proper attention to this phase. Example: normalising/checking names and other input.
- Data storage – Rushing from the input or output side to a database schema is probably an unwise way of proceeding. A more robust and stable development can be expected if an E-R or similar model is drawn up beforehand.
- Retrieval – proper recognition should be given to the obvious but neglected dependency of retrieval results on proper input. It is often erroneously presumed that quantity and quality of the input are sufficient for the expected searches. This is not unimportant in view of the interaction between certain work styles and search profiles.
- Formatting – In the OpenOffice study the discussion of citation formatting is a little bit biased: not every citation is indicated by 'a very short identifier' nor is always a reference list given. These are only typical for certain styles and certain disciplines. – Perhaps considering the output stage by itself would help giving some additional degree of freedom.

  In addition, it would have sufficed to point to packages like `natbib` by Patrick W. Daly to illustrate the varieties of formatting author names and the like. What is more interesting is the observation that the batch oriented formatting that e.g., BibTEX offers could be complemented by interactive uses, allowing better matching of citation text and context. (It would be interesting to try this, as it has implications for the workflow.)

2. The study by Job-Sluder et al. uses without discussion a relatively narrow understanding of 'bibliographical database' when it writes: 'a collection of bibliographical records'. If taken literally, not even quotes were admitted. This is certainly not intended, but serves to illustrate the dangers of taking certain concepts for granted.
3. Neither study touches upon those principled questions that a librarian for example might consider fundamental: what to describe, how to describe, and so on. While this might reflect to a degree the importance of the contributions that the library science has to offer, it maight also reflects a dangerous lack of awareness of possible problems and also opportunities. In the Knorr study for example, such questions were completely absent, although the author tried to give a comprehensive treatment in many other respects.

6 Conclusion

Both studies are welcome contributions which further a thorough understanding of user's needs and wishes.

For the future I would wish that

- more studies were based on observation, and
- more prior conceptualisation of the application domain

References

Rahmatollah Fattahi and Mehri Parirokh. Restructuring the bibliographic record for better organization and representation of knowledge in the global online environment. URL `http://www.um.ac.ir/~fattahi/ISKO/abstract1.htm`. Paper presented at the 7th ISKO International Conference, in Granada, Spain (10-13 July 2002) on Challenges in Knowledge Representation and Organization for the 21th Century: Integration of Knowledge across Boundaries„ 2002.

Kirk Job-Sluder, Greg Hanek, and Hong Zoang. User needs for bibliographic software. Technical report, OpenOffice.org, 2003. URL `http://php.indiana.edu/~csluder/OpenSource/`.

Dagmar Knorr. *Informationsmanagement für wissenschaftliche Textproduktionen*. 1998. ISBN 3-8233-5351-9.

Elaine Svenonius. *The Intellectual Foundation of Information Organization*. Digital Libraries and Electronic Publishing. MIT Press, Cambridge, Mass., 2000. ISBN 0-262-19433-3.