How can User-Oriented Depth Analysis be Constructively Guided?

Abstract: It is vital for library and information science to understand the subject indexing process thoroughly. However, document analysis, the first and most important step in indexing, has not received sufficient attention. As this is an exceptionally hard problem, we still do not dispose of a sound indexing theory. Therefore we have difficulties in teaching indexing and in explaining why a given subject representation is "better" than another. Technological advancements have not helped to close this fundamental gap. To proceed, we should ask the right questions instead. Several types of indexer inconsistencies can be explained as acceptable, yet different conceptualizations which resulting of the variety of groups dealing with a problem from their respective viewpoints. Multiple indexed documents are regarded as the normal case. Intersubjectively replicable indexing results are often questionable or do not constitute interesting cases of indexing at all. In the context of my ongoing dissertation in which I intend to develop an enhanced indexing theory by investigating improvements within a social sciences domain, this paper explains user-oriented selective depth analysis and why I chose that configuration. Strongly influenced by Mai's dissertation, I also communicate my first insights concerning current indexing theories. I agree that I cannot ignore epistemological stances and philosophical issues in language and meaning related to indexing and accept the openness of the interpretive nature of the indexing process. Although I present arguments against the employment of an indexing language as well, it is still indispensable in situations which demand easier access and control by devices. Despite the enormous difficulties the user-oriented and selective depth analysis poses, I argue that it is both feasible and useful if one achieves careful guidance of the possible interpretations. There is some hope because the number of useful interpretations is limited: Every summary is tailored to a purpose, audience and situation. Domain, discourse and social practice entail additional constraints. A pluralistic method mix that focuses on ecologically valid, holistic contexts and employs qualitative methods is recommended. Domain analysis urgently has to be made more practical and applicable. Only then we will be able to investigate empirically domains in order to identify their structures shaped by the corresponding discourse communities. We plan to represent the recognized problem structures and indexing questions of relevance to a small domain in formal, ontological computer models -- if we can find such stable knowledge structures. This would allow us to tailor dynamically summaries for user communities. For practical purposes we suggest to assume a less demanding position than Hjerland's "totality of the epistemological potential". It is sufficient that we identify and represent iteratively the information needs of today's user groups in interactive knowledge-based systems. The best way to formalize such knowledge gained about discourse communities is however unknown. Indexers should stay in direct contact with the community they serve or be part of it to ensure agreement with their viewpoints. Checklist/request-oriented indexing could be very helpful but it remains to be demonstrated how well it will be applicable in the social sciences. A frame-based representation or at least a sophisticated grouping of terms could help to express relational knowledge structures. There remains much work to do since in practice no one has shown yet how such an improved indexing system would work and if the indexing results were really "better".

1. Introduction: Why Subject Indexing is Hard and Why We still have No Sound Indexing Theory

We begin with the reiteration of common wisdom (cf. e.g., (Nohr, 1999)): Although subject indexing is at the very core of library and information science -- because a "useful" (conceptual) access structure by subject is an indispensable prerequisite to all retrieval --, we still lack a thorough understanding of this vital process: Which descriptions shall constitute
"good" subject representation(s) (subject access points) for a given document? Why? How can one arrive at them in a constructive and systematic way? Who determines the quality of a subject representation, and what accounts for a valid test?

What we need is a framework that explains why a certain conceptualization in indexing is more adequate than another and how we can find an even better suited one. Unfortunately, there is no such sound guidance. We do not dispose of a sufficiently explanatory or predictive theory that objectively or algorithmically can determine the subject(s) of a document. Although we do possess many indexing heuristics, and patents on indexing algorithms have been granted, their purported practical retrieval efficacy should not obscure the fact that there exists no sound theory behind. This is at least true from the humanistic point of view that indexing will always require the intellectual interpretation of a document's subject matter because only humans can understand the usage contexts in which the theoretically unlimited non-lexicalized paraphrases may constitute a concept (Fugmann, 1999).

We know that the first step in summarizing (subject analysis, in Mai's terminology: document analysis) is the most important. This is because it reduces most uncertainty and sets the scene for the following ones. However, it is also the least understood. For this very reason, most textbooks focus on later steps (e.g., the expression of subjects in certain indexing languages, i.e. the technical aspects of translating subject matter). Therefore, we should not be too surprised that in practice indexers are forced to work mostly on experience and intuition and that we have such big difficulties in teaching our students.

In fact the problem is so hard that it has effectively remained almost the same despite several decades of research. And one should not be tempted to view technological advancements as the remedy: The rapid proliferation of networked electronic documents and the increasing diversity of user groups and their heterogeneous needs has tightened the indexing bottleneck and at the same time led to a stronger demand for user- and depth-oriented document analysis.

2. Do Absolute Indexing Rules Exist and are Inconsistencies Always the Indexer's Fault?

We may have asked the wrong questions and need a more adequate conceptualization of the problem space. If there are no situation- and context-independent rules to discover we should instead utilize qualitative methods to investigate interpretation processes in small domains on per-case basis.

Often so-called "inconsistencies" are a natural outcome of different indexing contexts and purposes which lead to valid, multiple interpretations for the same document. There is no shortage of examples where the very same document was differently indexed by several indexing agencies who serve different user communities, and one cannot spot obvious "errors". Many of those inconsistencies are possible alternative conceptualizations that derive from the variety of groups dealing with a problem from their viewpoints. If we accept that a document can be (and typically is) multiply indexed and that different discourse communities need summaries tailored to their respective needs (Sigel, 1999 and 1997), the problem of subjectivity in indexing becomes less pressing. We "only" must find out how to keep all these contexts apart.

Furthermore, the difficulty of the indexing situation must be taken into account. Experience tells that indexing of certain formal document attributes is easier than indexing of abstract and compound subjects, and indexing of articles from "hard sciences", technology or technical documentation domains is easier than from the humanities and social sciences. We can more or less handle traditional document-oriented and summative indexing because these methods yield results which are easier to replicate, at least if one regards the conceptual level instead of the descriptor level. But the indexing cases on which even rather untrained students can reach a consensus are mostly uninteresting: Can we really speak of elaborated indexing when "the subject" of a document can be determined in an objectively and intersubjectively
consistent manner? We may want to automate this routine task and dedicate our time to more exciting problems.

3. What is and Why employ User-Oriented Depth Indexing in the Social Sciences?

In my ongoing dissertation work I intend to develop an enhanced indexing theory by investigating improvements within a social sciences domain (selected problems in migration). In this paper I explain user-oriented depth indexing and why I chose this configuration. In addition, I communicate first insights on current indexing theories. I am much indebted to Jens-Erik Mai whose dissertation on the Subject Indexing Process (Mai, 2000) has recently refocussed my thinking.

If my main purpose were to construct a retrieval system that performed well under today's conditions, I might be able to ignore (i.e., abstract out of my model) the heavy extra baggage of those epistemological stances and philosophical issues in language and meaning related to indexing. But such a system cannot help with the why? question. In order to make the right choices, I cannot simply disregard (inter alia) hermeneutics, linguistics and semiotics. We already know from Hjørlund (1997) that there exist several epistemological stances which are to a certain extent tied to specific indexing conceptions. Langridge (1989) and Mai (2000) have made it clear how deep indexing is rooted in language and meaning. Mai has also pointed out that one has to configure the right mixture in his continuum of 5 indexing conceptions according to the situation at hand.

I chose to focus on user-oriented depth indexing because I work at a specialized national documentation center whose task is to index the literature from the perspective of the clientele in selective depth. Thus I think rather in terms of communities than in terms of individual information needs. I am not sure if it is really necessary, recommendable or even feasible in this case to analyze individual information needs in order to guide this indexing process constructively. One may put too much emphasis on the often idiosyncratic needs, represent frequently changing short-term requirements, miss the bigger picture and eventually run out of resources. For a smaller center with more immediate contact to their users, requirement-oriented indexing could be an appropriate option. I am convinced that it is more fruitful to concentrate first (top-down) on the common needs that have emerged within a discourse community and that have been shaped by joint social practice and conventionalization. This does not exclude the perusal and generalization of selected individual requests.

As the name indicates, the user-oriented indexing conception is intentionally biased towards users and document usage. But it also involves (to a lesser extent) methods and aspects from content- and document-oriented, as well as from simplistic indexing. Different assumptions about the needs and prior knowledge of the target group will inevitably lead to different selections and conceptualizations. The representations may indeed become so extremely biased that they are no longer useful for other users at all.

While summative analysis can concentrate on the main, overall subject, exhaustive depth analysis (as best found in back-of-the-book indexes) would have to identify all mentioned concepts and themes. Database indexing in a specialized documentation center is typically more specific and closer to back-of-the-book indexing than in universal libraries. Selectivity means that because indexing is targeted towards user groups, the subjects of probable interest to them are indexed very exhaustively and specifically, but marginal ones are only touched and sometimes not even selected.

The user-oriented choice implies that:

- serving a group well is considered more important than the aim of neutrality and objectivity,
- we move from "the subject" to a collection of appropriate and useful concepts for a given user community.
As a consequence I cannot rely (as document-oriented indexing following rationalism does) on an objectively determinable subject matter. User-oriented indexing is connected to pragmatism, because it assigns concepts based on the estimation of the potential future uses of the document (in Hjørland's words: "the totality of a document's epistemological potential"). This assignment cannot be determined objectively and forever, because it is constantly influenced by ever changing discourse communities. The restrengthened rôle of the indexer also rules out automatic indexing: It would be congruent with Wittgenstein's thinking to ask for indexing programs to share the form of life and language game with the users! Instead interactive indexing tools are called for.

One may draw (like Mai) from C.S. Peirce's semiotics or recent text interpretation theories, or e.g. from Gadamer's hermeneutics: It seems convincing that the subject indexing process is necessarily of interpretive nature. It is not necessary that the document "contains" a subject which the indexer has to uncover. The indexer creates (constructs) the subjects via interpretation and expresses them in an indexing language. Let's therefore accept that indexing should be considered as an open, multi-step interpretive process with (theoretically) undetermined outcome and not as a determined function.

4. Do We need an Indexing Language at All?

Normally, the introduction of an indexing language is justified by the need to control certain retrieval parameters. This can be achieved by its devices. With interpretive openness and unlimited semiosis (the process of one sign generating the next sign and hence new interpretive levels), one might conclude that an indexing language would be more detrimental than helpful. In other words: Since natural language is so ambiguous, only natural language can convey the full meaning of natural language and ensure utility for future reference.

As the outcome of indexing can neither be prescribed nor controlled entirely (and hence good indexing cannot be guaranteed), we may even lose essential context because indexing is not per se homo-isomorph. Thus the usage of an indexing language could seriously distort the intended meaning. Considering the high costs of intellectual indexing, this may be the killer argument against intellectual interpretation before the actual retrieval, against indexing languages and indexers. I quote two examples:

- The simple conception of indexing keeps the error-prone process of interpretation completely out of the system. Humans are accustomed to learn each others interpretations (e.g. are familiar with an author's or indexer's form of life). They are clever enough to figure out meaning. Therefore meaning is best preserved if we leave the full text as it stands. The retrieval system just retrieves strings. It is the users who attribute meaning to them. Using just enough data will mutually disambiguate the words.
- In philosophical documentation, no normative thesaurus is possible. Therefore Henrichs (Henrichs, 1970) did not prescribe or pre-structure the vocabulary. He relied on keyword abstracts instead, using only words occurring in the text. However, he still used an indexing language because he determined and indicated by intellectual operation (and with numerical indices) which words form the thematic interrelations of text. This procedure shifts the burden and cost to the user who (at search time) has to navigate in the thesaurus-like word fields. Maintenance of the word fields was claimed to be manageable, but this may have changed over time.

Although it remains an open question, I argue that we still find domains where we clearly need an indexing language in order to express the numerous interpretations consistently. The reason is that for large repositories a precompiled optimal intellectual access structure is indispensable.
5. Why User-Oriented Depth Indexing is Difficult and may still be Both Feasible and Useful

Unfortunately, the choice of the highly "subjective" user-oriented depth indexing is faced with critical problems in indexing theory: Obviously, the potentially unlimited contexts within which a document could be used can assign far more subjects than we are used to from the (mentalistic) search for aboutness. What can keep me from assigning any number of subjects to a document and who is the arbiter against which I can check if a subject is permissible? If relevance and subject assignment lies in the requester: How do I get out of the relativity introduced by the users in their domains, their tasks, roles, functions and views? Is there any means to stop meaningless indexing?

The inherent vagueness in the indexing process makes predictions impossible and at best allows educated guesses. (But this situation does not seem to be much worse than e.g. in empirical retrieval tests). The need for epistemological openness (Kiel, 1994) makes it unfeasible to estimate reliably even part of the potential contribution a document could make to future information needs. We cannot predict all potential uses and we cannot prescribe which ideas and meanings would have a lasting value for the document (cf. (Mai, 2000, 239)).

I argue, however, that user-oriented indexing and selective depth analysis are both feasible and useful if one manages to guide the possible interpretations carefully. How can we achieve that in a principled way?

The interpretations are not arbitrary and not unrestrained as, fortunately, every summary is tailored to a purpose, audience and situation (Sparck Jones, 1999). Selective depth indexing can highlight those aspects of paramount importance for a certain community and can even ignore marginal others.

We need to investigate domains empirically to identify their structure which is shaped by discourse communities. Therefore, we urgently have to make domain analysis (Palmer, 1999; Hjørland & Albrechtsen, 1995; Albrechtsen, 1993) more practical and applicable. The results of the analysis (mostly: such entities related to problem structures or indexing questions which we consider worthwhile to add for their informational added value) are represented in formal, ontological computer models. We hope to find knowledge structures that are quite stable in a discourse over a certain time. Only then it is rewarding to model them. If we find them, we will be able to dynamically tailor summaries to user communities. In this sense, the stable element in knowledge organization is the cumulated information need of a group, while the flexible element is the adapted product.

To be able to develop operational procedures, we should assume a less demanding position than Hjørland's "totality of the epistemological potential". Since indexing is defined as the act of representing the interpretation(s) of a document for future use(s), it is not really necessary to find all interpretations ever possible, or to speculate on future information needs or document uses. As a first step it suffices:

- to identify and analyze empirically observed needs of today's users; and
- to employ modern information technologies to build interactive knowledge-based indexing tools.

One would constantly update and revise this knowledge base. But the best way to formalize such knowledge gained about discourse communities is hitherto unknown.

Fortunately, domain, discourse and social practice limit the number of interpretations useful in a certain discourse community. Direct contact to the community or being part of it ensures agreement with the viewpoints of the community the indexer serves.

Nevertheless, since indexing should be embedded in social practice, we run into typical problems of the social sciences (e.g. the observer's inevitable interaction with the observed, the dynamic and self-reflective nature of the process, the multitude of variables which one cannot control in ecological field settings and which, when controlled in laboratory settings, lead to unnatural observations or even artefacts).
Checklist indexing (fully developed in Soergel's description of request-oriented indexing (Soergel, 1985)) is an approach to systematically identify in a document such concepts of potential value to the users of an information system. This is accomplished by referring and checking against a recorded knowledge structure. To that purpose, the concepts gained from domain, text and request analysis typically are grouped into facets. The method has been successfully applied in certain domains. Since Mai (2000) argues that one may have difficulties in the humanities, it remains to be demonstrated how well it will be possible to transfer checklist/request-oriented indexing to a social sciences domain. But because the potential user group shapes the analysis of a document, an indexing tool would be of great help which could remind the indexer of such concept groups which certain user communities would like to have indexed.

6. Conclusion and Outlook

Subject indexing is very hard and we dispose of no sound explanative indexing theory. However, with a pluralistic method mix that focusses on ecologically valid contexts and that employs qualitative methods, user-oriented selective depth indexing should be both feasible and result in useful, not meaningless indexing.

In sum, what informs us about potential information needs of our users, if the subject matter cannot be determined looking at the document alone?

- More concrete knowledge about the user group and its activities. The same relativity introduced by the users' domains, their tasks, rôles, functions and views can alleviate the problem!

- Thorough wholistic domain analysis embedded in qualitative research will advance us.

- Investigating and understanding the cultural and social context in which the indexing result will be used will certainly help.

We do not know the best way to express and formalize knowledge gained about a discourse community. We envisage that a frame-based representation could turn out to be very useful manage relational knowledge structures (see e.g. Green 1997). In any case, even simple grouping of concepts into complex units should be advantageous compared to flat descriptor lists which cannot take arguments.

There remains much work to do since in practice no one has shown yet how such an improved indexing system would work and if it really resulted in "better" indexing (however to measure "goodness").

References


