Subject analysis and indexing: from automated indexing to domain analysis

Hanne Albrechtsen

Discusses the nature of subject analysis, suggesting three different conceptions of this: simplistic, content-oriented, and requirement-oriented, considering the type of subject information and indexing method appropriate for each.

Introduction

Preparing an index to a book or assigning indexing terms to documents involves the task of subject analysis. Subject analysis can be defined as 'the intellectual or automated process by which the subjects of a document are analyzed for subsequent expression in the form of subject data'.[1] Is automatic indexing a challenge, or even a threat to indexers? What do we mean when we talk about 'subjects' of books and other documents? Are there different conceptions of subjects, and hence of subject analysis; and if so, are such conceptions interconnected with methods applied for indexing?

These questions are important because discussions of indexing tend to focus on the performance of automated indexing[2] versus the performance of human indexing,[3] usually based on relevance judgments compiled from particular groups of IR-system users,[4] or observations of interindexer consistency.[5] Using performance as a criterion for finding a 'winning' approach to subject analysis and indexing implies a model which restricts itself to comparing the two techniques. Based on empirical observations of user and indexer behaviour, it represents a mechanistic evaluation method for subject assignment and retrieval. This article presents an alternative model for discussing subject analysis and indexing, where the choice between particular methods applied for this task is of minor importance. The intention is to attempt to place indexing in a wider social context beyond such mechanistic evaluation methods and to point towards new challenges for us as indexers.

Current methods for subject analysis and indexing

The process of creating subject data, i.e. subject entries for books or descriptors for documents in IR-systems, involves two major steps:

1. subject analysis of a document and expression of the perceived information in a concrete linguistic statement;
2. assigning the document with terms elicited in the linguistic statement, where these terms can be translated to conform with the terminology of a controlled vocabulary of indexing terms, for instance according to a thesaurus or a classification scheme.

This methodology is generally recommended as GIP (good indexing practice) to professional indexers. Examples include the International Standards Organization,[6] Foskett,[7] and Dutta & Sinha.[8] As pointed out by Frohmann[9] and by Blair,[10] the majority of the literature on subject indexing concentrates on step two and fails to provide precise rules for realizing step one where the challenge presented is: finding the subject(s) of a document. It should be mentioned, however, that the International Standards Organization recommends that the linguistic statement of subject contents follows a facetted approach to elicit concepts from a document. This latter method can be fruitful indeed, provided that it involves facets of a less generic nature than 'agent', 'instrument', 'object of action' etc.; compare for example ISO's general facets with Mutrux and Anderson's article[11] and Albrechtsen's facet scheme for software.[12] I shall return to such alternative approaches later.

What does it mean that a document has a subject?

After some years of magic sleep in the research community of library and information science, the concept of 'subject' has been awakened again to constitute a central area of study. Admittedly, the concept sometimes sneaked onto the stage for brief performances disguised under the term 'aboutness', as for example in domain analysis.[13] It has been argued[1] that the concept

of 'aboutness', introduced by Fairthorne and others,[14] was eventually introduced to avoid the difficulties of addressing the concept of 'subject' proper, and that the previous vagueness of the concept of 'subject' was transferred to the concept 'aboutness'. With a few exceptions, including Hutchins[15] and Beghtol,[13] both stressing the issue of intertextuality, the 'aboutness' approaches tended to handle documents as isolated sources of knowledge. At the same time, one may regard in particular Hutchins' work with discourse analysis, involving analysis of text organization as offering an interpretative approach to automatic text analysis, thus intending perhaps to inject a softer, humanistic flavour to what could otherwise be criticized as 'hard' automatic indexing approaches, based on statistical and/or computational linguistic techniques, approaches which have dominated research in indexing during the 1970s and 1980s, for example the work of Hutchins[15] and Salton.[16]

In contrast to the 'aboutness' approaches, at the same time challenging research in automatic indexing to reconsider their methodological weaknesses, Blair,[11] Hjørland,[17] Weinberg,[18] and Soergel[19] point towards new ways of looking at indexing. They reinstate the concept of 'subject' in the leading part for the practice and theory of indexing, stressing that the primary function that indexing should serve is the search for knowledge. They recommend that the indexer should not focus exclusively on the contents of documents, but attempt to anticipate the impact and value of a particular document for potential use. 'Indexing fails the researcher', says Weinberg,[18] because the indexes provide the 'aboutness' rather than the 'aspect', that is the point-of-view or innovation of a document, which cannot readily be perceived by interpreting a document as an isolated source of knowledge. Soergel[19] and Hjørland[17] both advocate that the traditional implicit definition of subjects as direct abstractions of single documents, which is also applied by researchers in automated indexing, as for example in the work of Korycinski and Newell,[2] is too narrow. Soergel[19] proposes a method of request-oriented indexing. Hjørland[17] offers a new definition of the concept of 'subject' as the totality of epistemological potentials of documents.

Conceptions of subject analysis

In order to have a clear reference frame for discussing subject analysis and indexing, I stipulate a model of conceptions of subject analysis and indexing. The model, shown in figure 1, covers three different conceptions or viewpoints of subject analysis and indexing.

The simplistic conception (i) regards subjects as absolute objective entities that can be derived as direct linguistic abstractions of documents or summed up like mathematical figures, using statistical indexing methods. According to this conception, indexing can be fully automated.

The content-oriented conception (ii) additionally involves an interpretation of the document contents that goes beyond the lexical and sometimes grammatical surface structure, which is the boundary within which the simplistic conception (i) operates. Subject

Conception of Subject Analysis and Indexing	Type of subject information	Indexing method
Simplistic Conception	Explicit information	Extraction
Content-oriented Conception	Implicit information	Assignment
Requirements-Oriented Conception		

Figure 1. Interconnection between conceptions of subject analysis, types of document information and indexing method.

analysis of document contents involves identification of topics or subjects that are not explicitly stated in the textual surface structure of a document, but they are readily perceived by a human indexer. Hence it involves a more indirect abstraction of the document itself.

The requirements-oriented conception (iii) regards subject data as instruments for transfer of knowledge, hence aiming at finding pragmatic information or knowledge. According to this conception, documents are created for communicating knowledge, and subject data should hence be tailored to function as instruments for mediating and rendering this knowledge visible to any possible interested persons.

I shall provide some critical comments on each approach, using examples from my work with classification and indexing of software and domain analysis of computing. From 1987 until 1991, I participated in the ESPRIT project PRACTITIONER, whose primary aim was to construct a thesaurus of software engineering terms, derived from automatic analysis of terminology applied in machine-readable software documentation.[20] While this approach was fruitful for getting an insight into terminological problems of the domain of software, it failed to facilitate a communication of knowledge about software beyond a narrow technical community. In contrast, domain analysis[12] regards software as involving both knowledge production and use in a broad societal and interdisciplinary context. Domain analysis entails investigations of scientific, sometimes technical, paradigms and viewpoints which represent different knowledge interests in one domain with the aim of building a classificatory structure to capture knowledge and serve its transfer and sharing. Even though discussions on possible drawbacks of approaches to automatic indexing and thesaurus construction already emerged during the work of the PRACTITIONER project,[21] the present model of subject analysis and indexing places itself in the context of this latter domain analysis approach to classification and indexing.

The simplistic conception of subject analysis

This conception regards subjects as direct abstractions of documents. Following this conception, to analyse and index software is equivalent to extracting automatically all single words or phrases from full-text software documentation.

It is often argued that technical documents like software documentation are amenable to fully automatic indexing approaches, due to a so-called 'hard' and unambiguous terminology applied by the document producers. This is not always true. For instance, experiments with automated and human indexing of UNIX online documentation[22] showed that the human indexers (software engineers) often perceived subjects in this documentation which were not readily stated in the text. One example: the UNIX utility COMM, which can be used for comparing two text files, does not feature the term 'compare' or any of its natural language equivalents in the machine-readable documentation. For automatic indexing of this software, this implies that users searching for utilities that can compare two files will not retrieve this item.

This is not to say that automatic techniques should be rejected as an approach to subject analysis. Rather, this has practical limitations. More seriously, one may contend that it lacks a theoretical foundation for subject analysis other than computational linguistics and statistics.

The content-oriented conception of subject analysis

The content-oriented conception of subject analysis is based on explicit as well as implicit subject information in texts. By explicit subject information is meant information which is expressed in the terminology applied by the document producer. A document may also convey implicit information, which is not directly expressed by the author, but which is readily understood or interpreted by a (human) reader of the document.

The content-oriented conception of subject indexing is the most common approach to subject indexing, including subject indexing of software.[23] However, the conception can be said to confine itself to representing or abstracting the document as an isolated entity. Consider for instance Frakes and Gandel's definition of subjects in software:

> 'A software representation is a mapping of predicates and terms describing individual objects and relationships among objects from the represented to the representing world. *The representing world may be an elaboration of the represented world, containing new predicates and terms*' (my italics).

And similarly, by Hall,[24] in the terminology of Saussure:

> '[The thesaurus terms] are signifiers and the [software] concept is the signified, with the preferred term acting as a surrogate for the concept . . .]

Such definitions of subjects in software reveal a content-oriented conception of subject indexing, where the descriptors (representations or signifiers) are predicates for the 'aboutness' or the intentional meaning of a document,[13] hence aiming at achieving an abstraction or a condensed version of software documentation.

The content-oriented conception of subject indexing implies some significant shortcomings, identified by Soergel[19] and Hjørland.[17] According to Soergel,[19] the content-oriented conception of subject analysis focuses on the document as an isolated source of knowledge, even though an indexer following this conception may consider the context of the document collection to which it belongs (intertextuality). As a result, the docu-

ment is assigned to one or more document classes or categories. Hjørland[17] argues that a pure content-oriented conception of subject indexing will often result in very trivial descriptors, which cannot be applied to search for more profound aspects like the theoretical reference frame applied, though often not stated in a document. This argument complies with Weinberg's critique of the 'aboutness' approaches to indexing.[18]

The requirements-oriented conception of subject analysis

The requirements-oriented conception of subject analysis is applied here as a common denominator for request-oriented approaches[19] and sociological-epistemological frameworks[17] for indexing. Subject analysis based on requirements entails a different focus from the content-oriented subject analysis approach. When analyzing a document, the indexer does not concentrate on representing or abstracting the explicit and implicit information in it. Rather, s/he asks: how should I make this document, or this particular part of it, visible to potential users? What terms should I use to convey its knowledge to those interested?

In requirements-oriented indexing it is the users' search for knowledge in IR-systems or indexes to books that determine the method for indexing. Hence, a document is analyzed with the purpose of predicting its potentials for serving particular groups of users. Soergel[19] advocates that these requirements should also determine the choice of indexing terms, and that the terminology applied by the indexer should comply with the terminology of the users. Soergel's approach places itself within the context of current developments in end-user searching; instead of searching for knowledge in an IR-system, supported by an intermediary, for instance a librarian, the user goes directly to the documents via a user-dedicated search vocabulary. This does not imply the extinction of intermediaries for mediating knowledge. Rather, Soergel's approach represents a change in the division of labour, in particular for the context of IR-systems, where the intermediary functions, traditionally performed by a reference librarian, are handed over to the indexers. However, for indexers of books, this intermediary role has always been a primary responsibility.

In order to achieve an anticipation of user needs, Soergel[19] proposes a bottom-up compilation of user terminology, for instance for the design of a user-dedicated in-house information system in the industry. This implies that the immediate requirements of a particular user group may be exaggerated at the expense of future users. What will happen when several inhouse databases featuring quite different terminologies and subject indexing approaches are to be integrated in the context of a large company take-over? Apart from such practical, though presumably not unsolvable, situations, one may wonder whether such empirical investigations of user terminology do eventually suffer from the same methodological deficiencies as automatic document indexing. The fact that user-dedicated vocabularies for particular IR-systems can be compiled automatically from logs of end-user search statements points in a mechanistic direction. One should hesitate, on such observations alone, to dismiss empirical investigations of user terminology for building requirements-oriented indexes or IR-systems. For special domains like fiction, fine arts and music, empirical investigations of user concepts, using for example association tests, may be a fruitful approach indeed, since these domains involve affective aspects in cognition, which cannot otherwise be captured.

Contrary to Soergel, but following his general idea, Hjørland proposes an approach of methodological collectivism for capturing knowledge interests or user needs.[1,17] He defines 'subject' as 'the totality of epistemological interests that one document may serve'.[17] Subjects are viewed as interconnected with scientific and human cognition. For subject analysis and indexing, this view imposes a priority on the indexer to decide on these long-term qualities of a document.

This theory presents a challenge for indexing practice, which I have pursued under the framework of domain analysis of computing.[13] Computing is a cross-disciplinary knowledge domain, which cannot be outlined readily on the basis of conventional criteria for the division of labour, or indeed of the sciences. Rather, it can be seen as a point of convergence or intersection between different interessees, including the producers as well as the users of computers. Based on methodological investigations of research and technical paradigms in connection with a critical appraisal of current methods for building classification systems and thesauri in this field, I proposed a faceted classification scheme with nine domain-specific facets and experimented with indexing different types of software applications. These facets can be used as a check-tag-list for subject analysis of software. Viewed as a structure, the check-tag-list is an indexing tool, complying with the faceted technique for subject analysis recommended by ISO.[6] Viewed as a framework for capturing concepts and aspects of computing, it constitutes an instrument for transfer of knowledge. In my present work as a teacher of indexing, I am introducing the students to domain and facet analysis of different disciplines in order to train them in building indexing tools which can serve several target groups for indexes. I am also investigating classificatory structures as interaction mechanisms in scientific communication.

Discussion and conclusion

Domain-specific approaches to designing indexing tools are open to critique. For instance, Horner[25] argues that at worst they suffer from the current trend towards fragmentation of knowledge rather than sup-

porting a holistic view. However, my application of domain analysis for indexing aims to facilitate transfer of knowledge in computing between different interessees. The objective is to investigate how domain-specific facets can be generalized between individual knowledge domains. In this context, I will mention a related approach to building indexing tools: the CIFT-project,[11] covering the domain of fine arts and literature and featuring aspect-oriented facets such as 'Scholarly approach' and 'Technique/method'. These facets are equivalent to the aspect-oriented facets 'Scientific Paradigm' and 'Technical approach' in my indexing framework for computing.

Domain analysis adheres to the requirements-oriented conception of subject analysis and indexing, which entails difficulties, when indexing practice enters the stage. Before dismissing the approach on suspicions of impracticality and, even worse, because it may involve the risks of isolating knowledge to an elite of domain experts, as implied in Horner's critique,[5] I will summarize the implications of the three different conceptions in the model presented in figure 1.

Subject analysis and indexing can be automated, but fully automated only according to a simplistic conception. The two other conceptions entail far more subjective and, indeed, difficult frameworks and methods for subject analysis. The difficulty lies in part in the fact that the concept of 'subject' still constitutes a pioneering area of study for research and practice in indexing.

All three conceptions have advantages and drawbacks. The advantage of following a simplistic conception is pragmatic: due to the current decrease in the pricing of computers and software, and the increasing cost of manpower, automatic indexing is hard to compete with on an immediate economical basis. The major drawback is that it does not aim at facilitating the transfer of knowledge to possible interessees in the documents it processes. This drawback may eventually render the technique expensive in the long run, if one considers the transfer and optimum use of knowledge as a key societal asset.

The content-oriented conception has the advantage of being an established technique for training and professional work in indexing, but together with the simplistic conception, it suffers from being one-sided, as it focuses on representing individual documents or document collections instead of considering their possible uses.

Finally, the requirements-oriented conception has the advantage that it supports the transfer and dissemination of knowledge. One may argue, however, that a major disadvantage for the practice of indexing is that its ultimate goal may only be realized in Utopia. How much 'scientific pre-thinking', following Soergel,[9] can we offer as professional indexers, and how do we train students of indexing to follow such a philosophy? How can indexers distinguish subjects of a high or low priority in a document and ensure its possible visibility in indexes and IR-systems for the future? How much responsibility should be imposed on us for judging or mediating the qualities of a document to potential users?

Instead of letting the answers to such questions blow in the wind, I suggest that indexers reconsider their practice. Current practice in indexing can be said to confine itself to modest, value-free ethics for dissemination of knowledge. Requirements-oriented indexing involves a high degree of subjectivity and responsibility in choosing among the qualities of documents. Current discussions in other professions, such as teaching and medical practice, tend to question prudent ethics of objectivity in mediating their services to their target groups. Rather than refraining from picking up the challenges posed by the social and cultural reality within which we operate, we should face the music, too. New frameworks like requirements-oriented approaches have potentials for supporting a broad and open transfer of knowledge, which is a primary responsibility of our profession. To choose such frameworks means the end of such 'threats' to our profession as automatic indexing, but more important, they provide us with a challenge to gain a new consciousness of the impact of our profession for mediating knowledge.

References

1. Hjørland, B. *Fundamental concepts in information science* (In Danish: Informationsvidenskabelige grundbegreber). Copenhagen: The Royal School of Librarianship, 1992, 40–3. [English edition in progress, co-authored by Hanne Albrechtsen].
2. Korycinski, C. and Newell, A.F. Natural-language processing and automatic indexing. *The Indexer* **17** (1) April 1990, 21–9.
3. Jones, K.P. Natural-language processing and automatic indexing: a reply. *The Indexer* **17** (2) October 1990, 114–15.
4. Cleverdon, C.W. Optimizing convenient on-line access to bibliographic databases. *Information Services and Use* 4 1984, 37–47.
5. Cooper, W.S. Is interindexer consistency a hobgoblin? *American Documentation* **20** (3) 1969, 266–78.
6. ISO [International Standards Organization]. *Methods for examining documents, determining their subjects and selecting indexing terms* 1985 (ISO 5963)
7. Foskett, A.C. *The subject approach to information*. London: Clive Bingley, 1982.
8. Dutta, S. and Sinha, P.K. Pragmatic approach to subject indexing—a new concept. *Journal of the American Society for Information Science* 35 1984, 325–31
9. Frohmann, B. Rules of Indexing: a critique of mentalism in information retrieval theory. *Journal of Documentation* 46 (2) 1990, 81–101.
10. Blair, D.C. *Language and representation in information retrieval*. Amsterdam: Elsevier Science Publishers. 1990.
11. Mutrux, R. and Anderson, J.D. Contextual indexing and faceted taxonomic access system. *Drexel Library Quarterly* 19 (3) Summer 1983, 91–111.

12. Albrechtsen, H. *Domain analysis for classification of software.* Copenhagen: The Royal School of Librarianship, 1992 (MSc Dissertation).
13. Beghtol, C. Bibliographic text theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation* 42 (2) 1986, 84–113.
14. Fairthorne, R.A. Content analysis, specification and control. *Annual Review of Information Science and Technology* 4 1969, 73–109.
15. Hutchins, J.W. On the problem of 'aboutness' in document analysis. *Journal of Informatics* 1 (1) 1977, 17–35.
16. Salton, G. and McGill, M. *Introduction to modern information retrieval.* New York: McGraw-Hill, 1983.
17. Hjørland, B. The concept of 'subject' in information science. *Journal of Documentation* 48 (2) 1992, 172–200.
18. Weinberg, B.H. Why indexing fails the researcher. *The Indexer* **16** (1) April 1988, 3–6
19. Soergel, D. *Organizing information—principles of database and retrieval systems.* New York: Academic Press, 1985.
20. Albrechtsen, H. Software concepts: knowledge organization and the human interface. *Advances in Knowledge Organization* edited by R. Fugmann. Frankfurt: Indeks Verlag, 1990, 48–64.
21. Albrechtsen, H. PRESS: A thesaurus-based information system for software reuse. *Classification research for knowledge representation and organization edited by N.J. Williamson and M. Hudon.* Amsterdam: Elsevier Science Publishers, 1992, 137–44.
22. Sedwell, I. and Albrechtsen, H. *The linguistic analysis of UNIX online documentation.* Uxbridge: Brunel University 1988 (ESPRIT PRACTITIONER project report, BrU-021).
23. Frakes, W.B. and Gandel, P.B. Representing reusable software. *Information and Software Technology.* 32 (10) 1990, 653–64.
24. Hall, P. *Domain modelling and concept storage.* Uxbridge: Brunel University 1990 (ESPRIT PRACTITIONER project report, BrU-IS-WP4.1–096).
25. Horner, D.S. Paradigms, discourses and language games: categorical frameworks and signs of the times. Presentation invited for expert panel on Domain Analysis in Information Science. *ASIS Annual Meeting, Columbus (OH) 22–28 October 1993.*

Hanne Albrechtsen is Lecturer in Classification and Indexing at the Royal School of Librarianship, Copenhagen, Denmark.

Justification for indexlessness?

Gillian Tindall prefaces her 20-page volume of literary criticism, *Rosamond Lehmann: an appreciation* (Chatto & Windus, 1985), with 'A note for the reader' concluding with this paragraph:

> This book has no formal index. Readers and reviewers are not, however, invited to take this as a sign of idleness or lack of scholarly attention on the part of myself or my editors: it has been a deliberate decision. The number of books published by Rosamond Lehmann is quite small, each is mentioned over and over again in my text and to most of them whole chapters are effectively devoted. This is not a biography; my text is not concerned with places, people or events so much as with themes. Any index which confined itself, in the traditional way, to concrete items, proper names, etc., would be too short and uninformative to be of much use; conversely one which attempted, by some means, to recapitulate the true nature of the book's contents would not make much sense except to someone who had already read the book. I have therefore preferred to set out the contents to each chapter rather more fully than is customary in the Contents, in the hope that this will be of sufficient assistance to any reader wishing to look up a specific book or phase of the author's life.

We print below also the contents to three chapters as set out in the list, and invite our readers' comments.

I THE PLACE YOU FIRST THOUGHT OF *page 7*
The author's basic themes – the location of her childhood – family – upbringing – education – *The Gipsy's Baby* – *The Swan in the Evening* – dead children – *A Note in Music* – 'The Red-Haired Miss Daintreys' – the sources of creation

II THE HOUSE ON THE HILL *page 24*
The house next door – *Dusty Answer* – the effects of the First World War – *Invitation to the Waltz* – The Desborough family – love doomed from the start

X THE INFERNAL GROVE *page 159*
The Echoing Grove: cycle of experience – perennial themes – mothers – sisters – betrayal – retrospective view – two different time-scales in this novel – attempts at an exact chronology – a cycle in the author's own life – triangular relationships – repetitions – this novel progressing thematically – the events of the story – the order in which they are recounted – cuff-links as symbolic objects – readers' reactions – theme of salvation – range of characters – sisters in opposition to one another – the outsider theme and its relationship to the writer herself

Our thanks to Gillian Tindall for permission to quote these passages.