

What's in a Name? Contextualizing the Document Concept¹

Helena Francke
Swedish School of Library and Information Science
University College of Borås/Göteborg University
helena.francke@hb.se

The term document is used in various contexts, often referring to very different things. This article argues that we need to avoid a restrictive, essentialist definition of the concept and instead study the cognitive models (Lakoff 1987a; 1987b) that guide our way of viewing documents in different situations. Examples are drawn from the Library and Information field to show how the view of documents is influenced by different cognitive models and how more complex understandings may be described in terms of clusters of models. Such a set of tools for discussing the concept will be particularly useful as we are facing a whole range of new types of 'documents' made possible by digital media.

Introduction

This article addresses the different meanings with which the term 'document' is understood in different disciplines and contexts, in particular in the two areas of Library and Information Science (LIS) and structured documents. While the notion of 'document' in both these areas is often considered secondary in importance to 'information' and

‘text,’ it has attracted quite some interest lately, not least because of the insistence on the document as a material object (Hansson *et al.*, 2003; Pédauque, 2003; cf. Hayles, 2002, who speaks of this in terms of the way materiality influences works as well as the reverse). An example is the discussion conducted in a French setting and reported in Roger T. Pédauque’s working paper ‘Document: Form, Sign and Medium, as Reformulated for Electronic Documents.’ (2003) It manifests an interest in documents that covers several disciplines and research areas, but also illustrates clearly that the understanding of the document concept is highly dependent on context. This serves to undermine the notion (to my mind questionable) of an essentialist approach to the concept, a critical standpoint that I will return to later.

The particular context in which I am working is characterized by an interest in document analysis, especially document structures and their relation to the material form of documents, as seen from the perspective of knowledge organisation, i.e. the representation of document collections for the purpose of storage, retrieval, and distribution of recorded knowledge. In my Ph.D. project, I try to integrate these aspects in an extended metaphor of Document Architecture, by which I propose an analytical document model used to highlight aspects of the individual document that are important to take into account when managing large document collections, particularly in digital media. In doing this, I am hijacking a concept which turns up once and again in computer discourse, often with different meanings. Sometimes it refers to the architecture in document managing software; sometimes it denotes the organisation of document components, with a primary focus on the text. What distinguishes my approach to document architecture compared to the general understanding of document architecture in computer science is that I look not only at the document structures and how they can be used by programs, but at how these structures relate to its social context, with regard to e.g. time (in terms of both production, preservation and dissemination), space, user interaction, and the interactions and relations between documents. This is where the connection with architecture comes into play; architecture viewed not merely as the building materials and how they have been joined, but also in the way the users of a

building interact with it, particularly in terms of space (cf. Zevi, 1993) and time.

Within literary studies, N. Katherine Hayles has pointed to the importance of including materiality as an aspect in the analysis of literary texts. Her focus is on how materiality, along with form, influences our experience of the text and acts together with the content to form the 'plot.' (Hayles, 2002)

As we move from literary analysis to dealing with documents in matters of storage, representation, preservation, distribution, or, for that matter, with presenting the literary text in a marked up version which reflects its fictional content and form, the material properties of the text are no longer of interest for the aspect they bring to the literary experience alone. The document serves a social function in our society in that it is assumed to provide us with knowledge and to preserve that knowledge for future generations.

The main focus in this article will be on the term 'document,' rather than on the architectural metaphor. What is meant by 'document' is, as we shall see, highly contextually determined in the two contexts from which I draw my main inspiration, and this is my attempt to find analytical instruments to discuss them.

Documents

The SGML standard definition of document reads 'A collection of information that is processed as a unit.' (ISO/DIS 8879-1986, p. 10) The ODA standard definition is slightly more specific: 'A structured amount of information intended for human perception, that can be interchanged as a unit between users and/or systems.' (ISO 8613:1, §4.27) This may perhaps be interpreted in physical terms as those signals in the computer memory that are treated as a unit, for instance since they are addressed via the same file name. In the SGML standard, the concept is closely related to that of 'SGML document,' which is ultimately not a physical construct but a logical one (ISO 8879-1986, p. 66), consisting of hierarchical document structures that divide the unit into a number of document elements. The type of elements and their reciprocal relations characterize a certain document type.

In Library and Information Science, documents have often been regarded as the physical form, or container, of information – what Michael Buckland has termed ‘information-as-thing’ (Buckland, 1991; cf. Smiraglia, 2001; Svenonius, 2000). This standpoint is sometimes an expression of the so-called conduit metaphor (Reddy, 1993), in which information is viewed, in rough terms, as a message that passes from sender to receiver through a medium – which may be a document – that is not affected by the message itself, nor affects the message. This uncomplicated view of information has been criticised (Day, 2000; 2001), and there are also more nuanced approaches, especially within the neighbouring area of bibliography, that pay attention to the materiality of documents, and point to the inseparability and interdependence of the message and the medium in which it is stored (McKenzie, 1999, pp. 12 f.; cf. Dahlström, 2000, 2002; Gunder, 2004; Hayles, 2002).

Another frequently used conception is of the document as a representation of a work, where the work is – in most cases – the prioritized concept (Lubetzky, 1953, cited in Smiraglia, 2001, p. 145; Gorman, 1980; IFLA, 1998; Smiraglia, 2001). This view finds a natural explanation in cataloguing practice. In C. A. Cutter’s often reproduced three principles for the functions of a library catalogue from 1904 (*Rules for a Dictionary Catalog*), for example, the first principle states that the catalogue should assist the user in identifying a document where the author, the title or the subject is known. In fact, these are all data that describe the work. The second principle, and part of the third, put focus on the actual *manifestations* of the work that are available in the library collection (Cutter, 1904, p. 12; cf. also IFLA, 1998, pp. 8 f.).

A different approach was made popular by a group of theorists and practitioners in early 20th Century Europe who have been clustered together under the name ‘documentalists,’ and whose ideas developed into information science. One of the consequences of their work was a renewed attempt to broaden the concept of document to other objects than those usually dealt with in libraries – such as books and manuscripts – and they suggested the practice of the documentalist in a ‘discursive and institutional system’ (Day, 2001, p. 38) as one basis for deciding what constituted a document. (cf. Day, 2001; Buckland, 1997) Basically, that which was being documented was a document. One of the most

ingenious contributors to the documentalist movement, Suzanne Briet, even included living creatures as documents, if they were the objects of a documentary act (Briet, 1951). A related view has been proposed by Niels Windfeld Lund, who emphasizes the activity involved in creating a document, so that the activity itself becomes a document, regardless of whether the result is a tangible object or not, such as a dance performance or a game of chess (Windfeld Lund, 2003; The Document Academy, 2003).

Birger Hjørland has suggested a more restrictive version of the institutional definition of document and proposed that a way of coming to terms with the problem of a very broad document concept is to restrict the objects of the documentary act to those that are documented by librarians; thus the meaning of the documentary act differs with the social or professional practice involved. This division of labour excludes from the field of responsibility of librarians for instance living creatures, such as Briet's famous antelope, which are the knowledge domain of the zoologist rather than of the librarian (Hjørland, 1995). Another way of singling out documents from the group of all informative objects is to focus on their function as evidence of human activity. In this case, messages of human activity are something that is contained by the documents and it is not human activity directed at the documents that make them such. Although not necessarily restricted to artefacts created or manufactured by humans, this position often puts emphasis on the social aspects of documents, not only in their ability to speak for us (Levy, 2002), but also in the way a shared experience of a document may create a spirit of community. (Levy, 2002; Brown and Duguid, 1996) It is my opinion that the stress on the social nature of the document and the implicit view of communication as something socially constructed and constructing means that this notion of a document does not end up with the uncomplicated view of the document—message relation of which the conduit metaphor has been criticized. Note, also, that these positions are not necessarily exclusive; some writers encompass two or several of them.

Two assumptions have, possibly more than any other, influenced the view of documents in LIS, and perhaps particularly in library practice. These are the different activities that have been performed with regard to documents, and the fact that these activities have to the greater part been

occupied with documents in the form of written text on paper, primarily books. Both these assumptions have been questioned with the introduction into library practice of huge amounts of digital material, which have brought to attention the impact of different storage and presentation media (Gunder, 2001), as well as multimedia forms, while at the same time making traditional library tasks a highly interesting area for people coming from other social practices, such as computer programming.

Prototype Effects and Fuzzy Categorization

As has already been hinted at, it is difficult to find an essence in the way that we use the concept of document today. The task has not become easier with the emergence of digital media. It has been pointed out that electronic documents have no materiality (although with Hayles we may question this assumption), which would make difficult the equation of 'document = material object' as a foundation for what a document is. Similarly, as soon as we move outside of a particular domain, it seems difficult to define 'document' purely in terms of institutional practices; who should have the preferential right of interpretation?

One way of approaching a core characteristic of document would be that it holds traces of human communication (Hjørland, 1995, p. 68; Levy, 2002). This in turn invites different ways of addressing issues concerned with documents. But there is one important drawback in viewing this criterion as *the* essence of 'document' rather than as *an* essential characteristic, namely that the problem of putting up boundaries to exclude that which is not a document remains. The problem of sharp boundaries that include or exclude objects or phenomena from categories have been discussed in terms of vagueness and fuzzy sets. Partly in this tradition, George Lakoff, in collaboration with Mark Johnson (Lakoff and Johnson, 1980; Lakoff, 1987a; Lakoff, 1987b), has proposed analytical instruments of categorization that are based on *prototype effects* and *interactional properties*.

Lakoff and Johnson (1980) argue, in a Wittgensteinian tradition, that concepts are defined in social interaction and in language use. Therefore, 'definition is not a matter of giving some fixed set of necessary and

sufficient conditions for the application of a concept [...]; instead, concepts are defined by prototypes and by types of relations to prototypes' (Lakoff and Johnson, 1980, p. 125). Prototype effects arise for different reasons. One such reason may be that there are *metonymic models* involved, so that what is perceived as a 'central case' is considered to stand for a whole group, as in the case of stereotypes. An example could be the way in which the printed text – and the codex in particular – for a long time has been viewed as the prototypical medium for a written text. This is also connected with a currently dominating *cognitive model* which is only beginning to be challenged with the introduction of different forms of digital carriers. To Lakoff, *idealised cognitive models* are used in theorizing about the world; they influence our attempts of making sense of the world (Lakoff, 1987b, p. 118; cf. McCauley, 1987). For example, a cognitive model embraced by those who use the Gregorian calendar states that there is a way of measuring time which is termed a week, and which consists of seven 24 hour periods. This, in turn, implies certain interactional properties, such as an agreement within a community that each 24 hour period contains a day and a night. Within LIS, a common cognitive model concerning documents is that a document has materiality and that this materiality in some way serves the purpose of preserving the content of the document. However, this does not necessarily imply recognition of the fact that the material form of the document influences its content; this may be considered a view governed by a different cognitive model (cf. new media studies such as Hayles, 2002). These cognitive models are not necessarily in conflict with each other, but could form *clusters* that allow the emergence of complex understandings (Lakoff, 1987a; 1987b). In the case of prototype effects, what Lakoff terms *representativeness structures* describe how we construct the relation between phenomena. But sometimes discursive categorizations include cases that are conventionalized variations on a central case, variations that cannot be 'predicted by general rules,' (Lakoff, 1987a, p. 75) that Lakoff calls *radial structures*. (Lakoff, 1987a; 1987b) Finally, in an interactional negotiation of meaning, we make use of certain properties that describe the types of relations or similarities that the different cases in a category share, namely *interactional properties*. These could, for example, have 'to

do with perception, motor activity, purpose, function, etc.’ (Lakoff and Johnson, 1980, p. 121)

In the discussion presented in Pédauque (2003) is suggested a taxonomy for digital document research. The research is grouped into three non-exclusive categories which are termed Form, Sign, and Medium. The first category of research is that which privileges the form, and the material and structural aspects of the document. Many researchers with this focus have an interest in documents in different medial forms, and Pédauque *et al.* attempt a preliminary description of the way documents in different media are understood. The print document is described as ‘*medium + inscription*’ (Pédauque, 2003, p. 5) and the electronic document as ‘*structure + data*’ (Pédauque, 2003, p. 6). The categories suggested are all fairly encompassing, and this one includes several of the ways of conceiving the document concept discussed above; certainly the view of documents within text encoding, but also to a large extent the way the document is viewed in the cataloguing process. However, an important part of providing access to documents is through subject representation, which falls within the next category – Sign – which considers the meaning of the document. Researchers in the last category – Medium – are most interested in the social role that the document plays, as ‘a trace, constructed or found, of a communication that exists outside space and time.’ (Pédauque, 2003, p. 3) This is where we may place some of the more sociological approaches to the document.

Each category is related to one aspect of a reading contract between a reader and a producer (Pédauque, 2003, p. 4). The document as Form is concerned with ‘an object of communication governed by more or less explicit formatting rules that materialize [the] reading contract’ (Pédauque, 2003, p. 4) which is interpreted in terms of legibility. The other two contracts are intelligibility and sociability. If we choose to look at clusters of cognitive models as a form of ‘cognitive supermodels’ – if such an expression may be allowed, because I believe that it is useful to work with hierarchies of cognitive models here as well – then I would propose that the three reading contracts outlined by Pédauque *et al.* may very well serve as such. The legibility contract thus implies for instance certain cognitive models having to do with the document as a structured entity with certain elements that are regular enough so that we may write rules

to describe them. This model fits well with such document types as scientific articles or technical documents. On the other hand, the intelligibility contract is of importance in understanding the view of documents as informative objects. The third reading contract – sociability – concerns for instance the document as evidence of human activity, and the sociology of documents.

To give an example of how these different cognitive models may come into play in the interaction with a document, one may note that a librarian cataloguing a document in the local OPAC may well be moving between all three of these cognitive models: the physical properties of the document, such as height, width, number of pages, etc. are entered, along with intellectual properties having to do with the production and history of the document (author, production year, owner), descriptors describing the document's intellectual content (keywords, classification code), and possibly also a suggestion for the user group that the document is most suitable for (school children, general public, experts). If the document is part of a citation service, the document may also be positioned socially in that it is integrated in a system which records the network of references in which it takes part. I would suggest that cataloguers are then engaging a cluster of cognitive models in their work, a cluster that researchers need to be aware of when studying the cataloguers' practice.

Conclusion

The basis of the notion of document in Document Architecture is to be found in the cognitive model of the legibility contract, both in Document Architecture's traditional uses and in the sense proposed here. However, there are aspects from the other reading contracts that are interesting to highlight in an analysis. If we want to emphasize that material and structural form influence meaning, then the intelligibility contract cannot be ignored. Similarly, if some form of recorded human communication is what characterizes the document, then we are talking sociability.

Exactly how the distribution of the different cognitive models will look, and which models will cluster to inform the understanding of documents, I believe is dependent on the purpose for which the concept

is being used. Rather than trying to find the ultimate definition, for other than lexicographical purposes, each study will need to identify the interactional properties that are of interest in that particular case, as well as be aware of which prototypical effects are at work. It is my belief that thinking in these terms may facilitate a self-conscious use of 'document,' and in turn, that thinking in terms of documents may influence the way we work with such intangible entities as 'information' or 'text.'

Notes

1. The article is based on a presentation made at the ALLC/ACH 2004, The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers in the Humanities, Göteborg University 11-16 June, 2004.

References

- Briet, S.** (2003). What is Documentation? Transl. R. E. Day and L. Martinet. [*Qu'est-ce que la documentation?* Paris: Éditions Documentaires Industrielles et Techniques, 1951.] Detroit: Wayne State University. <http://www.lisp.wayne.edu/~ai2398/briet.htm>, (accessed 2 June 2004).
- Brown, J. S. and Duguid P.** (1996). The Social Life of Documents, *First Monday*, 1(1). <http://www.firstmonday.org/issues/issue1/documents>, (accessed 2 June 2004).
- Buckland, M.** (1991). *Information and Information Systems*. Westport, CT: Praeger.
- Buckland, M. K.** (1997). What is a 'Document'? *Journal of the American Society for Information Science* 48(9): 804-809.
- Cutter, C. A.** (1904). *Rules for a Dictionary Catalog*. 4th ed. Washington, DC: Government Printing Office.
- Dahlström, M.** (2000). Drowning by Versions, *Human IT* 4(4): 7-38. <http://www.hb.se/bhs/ith/4-00/md.htm>, (accessed 2 June 2004).
- Dahlström, M.** (2002). Nya medier, gamla verktyg [New Media, Old Tools], *Human IT* 6(4): 71-116. <http://www.hb.se/bhs/ith/4-02/md.pdf>, (accessed 2 June 2004).
- Day, R. E.** (2000). The 'Conduit Metaphor' and the Nature and Politics of Information Studies, *Journal of the American Society for Information Science* 51(9): 805-811.
- Day, R. E.** (2001). *The Modern Invention of Information: Discourse, History, and Power*. Carbondale, IL: Southern Illinois University Press.
- The Document Academy** (2003). Studieprogram DOKVIT De fire bilder [Study Program DOKVIT The Four Images]. Institutt for dokumentasjonsvitenskap, Det humanistiske fakultet, Universitetet i Tromsø. <http://thedocumentacademy.hum.uit.no/studieprogram/index.html>, (accessed 2 June 2004).
- Gorman, M.** (1980). AACR2 Main Themes. In Hargrett Clack, D. (ed), *Making of a Code: The Issues Underlying AACR2. Papers given at The International Conference on AACR2 held March 11-14, 1979 in Tallahassee, Florida*. Chicago: American Library Association, pp. 41-50.
- Gunder, A.** (2001). Forming the Text, Performing the Work: Aspects of Media, Navigation, and Linking, *Human IT* 5(2-3): 81-206. <http://www.hb.se/bhs/ith/23-01/ag.htm>, (accessed 8 June 2004).

- Gunder, A.** (2004). Harry Ludens: *Harry Potter and the Philosopher's Stone* as a Novel and Computer Game, *Human IT* 7(2): 1-137. <http://www.hb.se/bhs/ith/2-7/ag.pdf> (accessed 22 Dec. 2004).
- Hansson, J., Francke, H., Dahlström, M., and Gunnarsson, M.** (2003). Documents in Library and Information Science: Sociotechnical Dimensions in Document Genre and Architecture Studies. Paper presented at *The Annual Meeting of the Document Academy*, Berkeley, CA, 13-15 August 2003. <http://thedocumentacademy.hum.uit.no/events/docam03.abstracts/boras.paper.html>, (accessed 2 June 2004).
- Hayles, N. K.** (2002). *Writing Machines*. (Mediawork Pamphlet Series). Cambridge, MA: MIT Press.
- Hjørland, B.** (1995). *Informationsvidenskabelige grundbegreber* [Fundamental Concepts in Information Science]. Vol. 1. 2nd rev. ed. København: Danmarks biblioteksskole.
- IFLA/International Federation of Library Associations and Institutions, Study Group on the Functional Requirements for Bibliographic Records** (1998). *Functional Requirements for Bibliographic Records*. (UBCIM Publications; New Series Vol. 19). München: K. G. Saur.
- ISO/DIS 8613/1-6** (1986). *Information Processing – Text and Office Systems – Office Document Architecture (ODA) and Interchange Format*.
- ISO 8879-1986** (1986). *Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*. First ed.
- Lakoff, G.** (1987a). Cognitive Models and Prototype Theory. In Neisser, U. (ed), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. (Emory Symposia in Cognition; 1). Cambridge: Cambridge University Press, pp. 63-100.
- Lakoff, G.** (1987b). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL and London: University of Chicago Press.
- Lakoff, G. and Johnson, M.** (1980). *Metaphors We Live By*. Chicago, IL and London: University of Chicago Press.
- Levy, D. M.** (2002). Documents and Libraries: A Sociotechnical Perspective. In Peterson Bishop, A., Van House, N. A., and Battenfield, B. P. (eds), *Digital Library Use: Social Practice in Design and Evaluation*. Cambridge, MA: MIT Press, pp. 25-42.
- McCauley, R. N.** (1987). The Role of Theories in a Theory of Concepts. In Neisser, U. (ed), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. (Emory Symposia in Cognition; 1). Cambridge: Cambridge University Press, pp. 288-309.
- McKenzie, D. F.** (1999). *Bibliography and the Sociology of Texts*. Cambridge: Cambridge University Press.
- Pédauque, R. T.** (2003). Document: Form, Sign and Medium, As Reformulated for Electronic Documents. Working paper, version 3, July 8. STIC-CNRS. http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/94/sic_00000594_01/sic_00000594.pdf, (accessed 2 June 2004).

Helena Francke, Preprint of article in *Literary and Linguistic Computing* 2005 20(1):61-69.

Reddy, M. J. (1993). The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language. In Ortony, A. (ed), *Metaphor and Thought*. 2nd ed. New York: Cambridge University Press, pp. 164-201.

Smiraglia, R. P. (2001). *The Nature of 'A Work:' Implications for the Organization of Knowledge*. Lanham, MD and London: Scarecrow Press.

Svenonius, E. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, MA: MIT Press.

Windfeld Lund, N. (2003). Doceo + Mentum = Document: A Ground for a New Discipline? Paper presented at *The Annual Meeting of the Document Academy*, Berkeley, CA, 13-15 August 2003. <http://thedocumentacademy.hum.uit.no/events/docam/03/abstracts/lund.paper.html>, (accessed 2 June 2004).

Zevi, B. (1993). *Architecture as Space: How to Look at Architecture*. Rev. ed. New York: Da Capo Press.