What is a "digital document"?

Michael Buckland
School of Information Management and Systems,
University of California, Berkeley, CA, USA 94720-4600
buckland@sims.berkeley.edu

Abstract The question "What is a digital document?" is seen as a special case of the question "What is a document?" Ordinarily the word "document" denotes a textual record. Early this century, attempts to provide access to the rapidly growing quantity of available documents raised questions about which should be considered a "document". Paul Otlet and others developed a functional view of "document" and discussed whether, for example, sculpture, museum objects, and live animals, could be considered to be "documents". Suzanne Briet equated "document" with organized physical evidence. These ideas resemble notions of "material culture" in cultural anthropology and "object-as-sign" in semiotics. Others, especially in the USA (e.g. Jesse Shera and Louis Shores) took a narrower view. Old confusions between medium, message, and meaning are renewed with digital technology because technological definitions of "document" become even less realistic when everything is in bits.

1. Introduction

When we refer to a paper document, a papyrus document, or a microfilmed document, the meaning is clear. However, the idea of a "digital document" is more difficult. We can recognize e-mail and a technical report generated by a wordprocessor as digital documents, but beyond these simple examples the concept of a "document" becomes less clear. Is a software program a document? It has lines of language-like text. Is an operating system a document? One can enumerate different types of digital documents and this is necessary because of the need to specify standards in order to achieve efficiency and interoperability. But if one seeks completeness, the process becomes arbitrary and intellectually unsatisfying because it is not clear where the frontier between documents and non-documents should be.

A paper document is distinguished, in part, by the fact that it is on paper. But that aspect, the technological medium, is less helpful with digital documents. An e-mail message and a technical report exist physically in a digital environment as a string of bits, but so does everything else in a physical environment. "Multimedia," which used to denote multiple, physically-different media, is now of renewed interest, because, ironically, the multiple media can be reduced to the mono-medium of electronically stored bits.

For practical purposes, people develop pragmatic definitions, such as "anything that can be given a file name and stored on electronic media" or "a collection of data plus properties of that data that a user chooses to refer to as a logical unit." And, as so often in discussions of information, one finds definitions of "document" that focus on one aspect and are often highly metaphorical, such as "`captured' knowledge," "data in

context," and "an organized view of information."

Digital systems have been concerned primarily with text and text-like records (e.g. names, numbers, and alphanumeric codes), but the present interest in icons and graphics reminds us that we may need to deal with any phenomena that someone may wish to observe: events, processes, images, and objects as well as texts [BUC 91].

2. From document to "documentation"

Digital documents are relatively new, but discussion of the broader question "What is a document?" is not new. In the late 19th century there was increasing concern with the rapid increase in the number of publications, especially of scientific and technical literature and of social "facts." Continued effectiveness in the creation, dissemination, utilization of recorded knowledge was seen as a needing new techniques for managing the rising flood of documents.

The "managing" that was needed had several aspects. Efficient and reliable techniques were needed for collecting, preserving, organizing (arranging), representing (describing), selecting (retrieving), reproducing (copying), and disseminating documents. The traditional term for this activity was "bibliography". However, "bibliography" was not entirely satisfactory. It was felt that something more than traditional "bibliography" was needed, e.g. techniques for reproducing documents and "bibliography" also had other well-established meanings in related to traditional techniques of book-production.

Early in the 20th century the word "documentation" was increasingly adopted in Europe instead of "bibliography" to denote the set of techniques needed to manage this explosion of documents. From about 1920 "documentation" (and related words in English, French and German) was increasingly accepted as a general term to encompass bibliography, scholarly information services, records management, and archival work. After about 1950 more elaborate terminology, such as "information science", "information storage and retrieval", and "information management", increasing replaced the word "documentation".

3. From documentation back to "document"

The problems created by the increase in printed documents did lead to the development of techniques developed to manage significant (or potentially significant) documents, meaning, in practice, printed texts. But there was (and is) no theoretical reason why documentation should be limited to texts, let alone printed texts. There are many other kinds of signifying objects in addition to printed texts. And if documentation can deal with texts that are not printed, could it not also deal with documents that are *not* texts at all? How extensively could documentation be applied? Stated differently, if the term "document" were used in a specialized meaning as the technical term to denote the objects to which the techniques of documentation could be applied, how far could the scope of documentation be extended. What could (or could not) be a document? However, the question was not often formulated in these terms.

An early development was to extend the notion of document beyond written texts, a usage to be found in major English and French dictionaries. "Any expression of human thought" was a frequently used definition of "document" among documentalists. In the USA, the phrases "the graphic record" and "the generic book" were widely used. This was convenient for extending the scope of the field to include pictures and other graphic and audio-visual materials. The Belgian Paul Otlet (1868-1944) is known for his observation that documents could be three dimensional, thereby including sculpture. From 1928, museum objects were likely to be included by documentalists within definitions of "document" (e.g. DUP 33). The overwhelming practical concern of documentalists was with printed documents, so the question of how far the definition of "document" could be extended received little direct attention. Occasionally a thoughtful writer would discuss the topic, perhaps because interested in some novel form of signifying object, such as educational toys, or because of a desire to theorize.

Paul Otlet: Objects as documents

Otlet extended the definition of "document" half-way through his *Traité de documentation* of 1934 [OTL 34]. Graphic and written records are representations of ideas or of objects, he wrote, but the *objects themselves* can be regarded as "documents" if you are informed by observation of them. As examples of such "documents" Otlet cites natural objects, artifacts, objects bearing traces of human activity (such as archaeological finds), explanatory models, educational games, and works of art (OTL 34: p. 217]; also [OTL 90: pp. 153 & 197], and [IZQ 95]).

In 1935 Walter Schuermeyer wrote: "Nowadays one understands as a document any material basis for extending our knowledge which is available for study or comparison." ("Man versteht heute unter einem Dokument jede materielle Unterlage zur Erweiterung unserer Kenntnisse, die einem Studium oder Vergleich zugänglich ist." [SCH 35: p. 537]).

Similarly, the International Institute for Intellectual Cooperation, an agency of the League of Nations, working in collaboration with Union Français des Organismes de Documentation, developed technical definitions of "document" and related technical terms in English, French and German versions:

"Document : Toute base de connaissance, fixée matériellement, susceptible d'être utilisée pour consultation, étude ou preuve. Exemples: manuscrits, imprimés, représentations graphiques ou figurés, objets de collections, etc...

Document : Any source of information, in material form, capable of being used for reference or study or as an authority. Examples : manuscripts, printed matter, illustrations, diagrams, museum specimens, etc.... ([ANO 37: p. 234])

Suzanne Briet: Physical evidence as document

Suzanne Briet (1894-1989), the perceptive French librarian, addressed the extension of the meaning of "document" with unusual directness. (For Briet, also known as Suzanne Dupuy and as Suzanne Dupuy-Briet, see [LEM 89], [BUC 95], [BUC 97b]). In 1951 Briet published a manifesto on the nature of documentation, *Qu'est-ce que la documentation*, which starts with the assertion that "A document is evidence in support of a fact." ("Un document est une preuve à l'appui d'un fait" ([BRI 51: p. 7]). She then elaborates: A document is "any physical or symbolic sign, preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon". ("Tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène ou physique ou intellectuel." p. 7.) The implication is that one should consider documentation to be concerned with access to evidence rather than with access to texts.

Briet gives examples: A star in the sky is not a document, but a photograph of it would be; a stone in a river is not a document, but a stone exhibited in a museum would be; an animal in the wild is not a document, but a wild animal presented in a zoo would be. An antelope running wild on the plains of Africa should not be considered a document, she rules. But if it were to be captured, taken to a zoo and made an object of study, it has been made into a document. It has become physical evidence being used by those who study it. Indeed, scholarly articles written about the antelope are secondary documents, since the antelope itself is the primary document.

Briet's rules for determining when an object has become a document are not made clear, but her discussion seems to indicate that:

1. There is materiality: Only physical objects can be documents, cf. [BUC 91];

2. There is intentionality: It is intended that the object be treated as evidence;

3. The objects have to be processed: They have to be made into documents; and, we think,

4. There is a phenomenological position: The object is perceived to be a document.

This situation is reminiscent of discussions of how an image is made art by framing it as art. Did Briet mean that just as "art" is made art by "framing" (i.e. treating) it as art, so an object becomes a "document" when it is treated as a document, i.e. as a physical or symbolic sign, preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon?

Ron Day ([DAY 96]) has suggested, very plausibly, that Briet's use of the word "indice" is important, that it is indexicality--the quality of having been placed in an organized, meaningful relationship with other evidence--that gives an object its documentary status.

Donker Duyvis: A spiritual dimension to documents

Frits Donker Duyvis (1894-1961), who succeeded Paul Otlet as the central figure in the International Federation for Documentation, epitomized the technological modernism of the documentalists in his dedication to the trinity of scientific management, standardization, and bibliographic control as complementary and mutually reinforcing bases for achieving progress ([ANO 64]). Yet Donker Duyvis was not a materialist. He adopted Otlet's view that a document was an expression of human thought, but he did so in terms of Anthroposophy, a spiritual movement based on the notion that there is a spiritual world comprehensible to pure thought and accessible only to the highest faculties of mental knowledge. As a result, Donker Duyvis was sensitive to what we might now call the cognitive aspects of the medium of the message. He wrote that:

"A document is the repository of an expressed thought. Consequently its contents have a spiritual character. The danger that blunt unification of the outer form exercises a repercussion on the contents in making the latter characterless and impersonal, is not illusory.... In standardizing the form and layout of documents it is necessary to restrict this activity to that which does not affect the spiritual contents and which serves to remove a really irrational variety." ([DON 42]. Translation from [VOO 64: p. 48])

Ranganathan: Micro-thought on a flat surface

The Indian theorist S. R. Ranganathan, usually so metaphysical, took a curiously narrow and pragmatic position on the definition of "document", resisting even the inclusion of audiovisual materials such as radio and television communications. "But they are not documents; because they are not records on materials fit for handling or preservation. Statues, pieces of china, and the material exhibits in a museum were mentioned because they convey thought expressed in some way. But none of these is a document, since it is not a record on a more or less flat surface." (RAN 63: p. 41]).

Ranganathan's view of "document" as a synonym for "embodied micro thought" on paper "or other material, fit for physical handling, transport across space, and preservation through time" was adopted by the Indian Standards Institution ([IND 63: p. 24]). Others, also, took a limited view of what documents were. In the USA, two highly influential authors opted for a view of documents that was only an extension of textual records to include audiovisual communications. Louis Shores popularized the phrase "the generic book" (e.g. [SHO 77]) and Jesse H. Shera used "the graphic record" with much the same meaning (e.g. [SHE 72]).

4. Anthropology: Material culture

Otlet was explicit that his view of "document" included archaeological finds, traces of human activity, and other objects not intended as communication. "Collections of objects brought together for purposes of preservation, science and education are essentially documentary in character (Museums and Cabinets, collections of models,

specimens and samples). These collections are created from items occurring in nature rather than being delineated or described in words; they are three dimensional documents." ([OTL 20]. Translation from [OTL 90: p. 197]).

The notion of objects as documents resembles the notion of "material culture" among cultural anthropologists "for whom artifacts contributed important evidence in the documentation and interpretation of the American experience." ([AME 85: p. ix) and in museology (e.g. [KAP 94], [PEA 90]).

5. Semiotics: "Text" and "object-as-sign"

Briet's ideas concerning the nature of a "document" invite discussion in relation to semiotics. In this context we note Dufrenne's discussion of the distinction between aesthetic objects and signifying objects:

"The function of such [signifying] objects is not to subserve some action or to satisfy some need but to dispense knowledge. We can, of course, call all objects signifying in some sense. However, we must single out those objects which do more than signify merely in order to prepare us for some action and which are not used up merely in the fulfillment of the task. Scientific texts, catechisms, photograph albums, and, on a more modest scale, signposts are all signs whose signification engages us in an activity only after having first furnished us with information." ([DUF 73: p. 114]).

We can observe that by the inclusion of museum and other "found" objects, Briet's "any physical or symbolic sign" appears to include both human signs and natural signs. Others developed the notion of "object-as-sign". Roland Barthes, for example, in discussing "the semantics of the object", wrote that objects "function as the vehicle of meaning: in other words, the object effectively serves some purpose, but it also serves to communicate information: we might sum it up by saying that there is always a meaning which overflows the object's use." ([BAR 88: p. 182]). We can note the widespread use of the word "text" to characterize patterns of social phenomena not made of words or numerals, but there seems to have been relatively little attention to the overlap between semiotics and information management. (See, however, [WAR 90].)

One difference between the views of the documentalists discussed above and contemporary views is the emphasis that would now be placed on the social construction of meaning, on the viewer's perception of the significance and evidential character of documents. In semiotic terminology,

"...signs are never natural objects... The reason is simply that the property of being a sign is not a natural property that can be searched for and found, but a property that is given to objects, be they natural or artificial, through the kind of use that is made of them. Both as objects and as means, signs have to be treated as something invented, and in this sense they are correlated to actions." ([SEB 94: v. 1, p. 18]).

Briet's notion of documents as evidence can occur in at least two ways. One purpose of

information systems is to store and maintain access to whatever evidence has been cited as evidence of some assertion. Another approach is for the person in a position to organize artefacts, samples, specimens, texts, or other objects to consider what it could tell one about the world that produced it, and then, having developed some theory of its significance to place the object *in evidence*, to offer it as evidence by the way it is arranged, indexed or presented. In this manner information systems can be used not only in finding material that already is in evidence, but also in arranging material so that someone may be able to make use of it as (new) evidence for some purpose. ([WIL 95]).

6. Digital documents

The evolving notion of "document" among Otlet, Briet, Schuermeyer, and the other documentalists increasingly emphasized whatever functioned as a document rather than traditional physical forms of documents. The shift to digital technology seems to make this distinction even more important. Levy's thoughtful analyses have shown that an emphasis on the technology of digital documents has impeded our understanding of digital documents as documents (e.g. [LEV 94]). Every thing in digital technology is stored as a string of bits, so the usual physical form (on paper, on microfilm) no longer helps. In this sense, any distinctiveness of a document as a physical form is further diminished.

Fifty years ago, one would look up logarithmic values in a printed book of "log tables" in order to do calculations. The volume of log tables was a conventional document. Today, one could imagine using a set of log tables stored online, which could be regarded as a digital version of the printed log tables. However, it is more likely that one would use an algorithm to compute log values as needed. The answer given should be the same. Perhaps one does not know whether the computer has used a table or an algorithm. The table and the algorithm seem functionally equivalent. What has happened to the notion of a "document"? One answer is that whatever is displayed on the screen or printed out is a document. One might say that the algorithm is functioning as a document, as a dynamic kind of document, one that reminds us of Otlet's view that an educational toy should be considered to be a kind of document. It would be consistent with the trend, described above, towards a defining a document in terms of function rather than physical format.

Each different technology has different capabilities, different constraints. If we sustain the functional view of what constitutes a document, we should expect documents to take different forms in the contexts of different technologies and so we should expect the range of what could be considered a document to be different in a digital and paper environments. The algorithm for generating logarithms, like a mechanical educational toy, can be seen as a dynamic kind of document unlike ordinary paper documents, but still consistent with the etymological origins of "docu-ment", a means of teaching - or, in effect, evidence, something from which one learns.

Attempts to define digital documents are likely to remain elusive, if more than an ad hoc, pragmatic definition is wanted. Definitions based on form, format and medium appear to

be less satisfactory that a functional approach, following the path of reasoning underlying the largely forgotten discussions of Otlet's objects and Briet's antelope.

[AME 85] Ames, K. L. et al. *Material culture: a research guide*, ed. by T. J. Schlereth. University Press of Kansas, Lawrence, Kansas, 1985.

[ANO 37] Anon. La terminologie de la documentation, *Coopération Intellectuelle*, 77, pp. 228-240, 1937.

[ANO 64] Anon. *F. Donker Duyvis: His life and work*, (NIDER publ. ser. 2, no. 45), Netherlands Institute for Documentation and Filing, The Hague, Netherlands, pp. 39-50, 1964.

[BAR 88] Barthes, R. *The semiotic challenge*. Hall & Wang, New York, 1988.

[BRI 51] Briet, S. *Qu'est-ce que la documentation*. EDIT, Paris, 1951.

[BUC 91] Buckland, M. K. Information as thing. *Journal of the American Society of Information Science* v, 42, pp. 351-360, 1991.

[BUC 95] Buckland, M. K. The centenary of `Madame Documentation': Suzanne Briet, 1894-1989, *Journal of the American Society for Information Science*, 42, pp. 586-588, 1995.

[BUC 97a] Buckland, M. K. Suzanne Briet, 1894-1989. In: *Dictionnaire encyclopédique de l'information et de la documentation*. (Collection REF). Editions Nathan, Paris, pp. 105-106, 1997.

[BUC 97b] Buckland, M. K. What is a "document"?, *Journal of the American Society for Information Science* 48, pp. 804-809, 1997.

[DAY 96] Day, Ron. Private communication, 1996.

[DON 42] Donker Duyvis, F. *Normalisatie op het gebied der documentatie*. [Standardization in the domain of documentation]. (NIDER publ. 214). NIDER, The Hague, Netherlands, 1942.

[DUF 73] Dufrenne, M. *The phenomenology of aesthetic experience*, Northwestern University Press, Evanston, Illinois, 1973.

[DUP 33] Dupuy-Briet, S. Rapport présenté à la Commission de terminologie. In: *International Institute for Documentation. XIIe Conférence. Rapport*. Bruxelles, 1933. (IID publication 172a), pp. 187-192, IID: Brussels, 1933.

[FRA 78] Frank, P. R. *Von der systematischen Bibliographie zur Dokumentation*. (Wege der Forschung 144). Wissenschafliche Buchgesellschaft, Darmstadt, 1978.

[IND 63] Indian Standards Institute. *Indian standard glossary of classification terms. IS : 2550 - 1963*, Indian Standards Institute, New Delhi, 1963.

[IZQ 95] Izquierdo Arroyo, J. M. *La organizacion documental del conocimiento*. Tecnidoc, Madrid, 1995.

[KAP 94] Kaplan, F. E. S., ed. *Museums and the making of "ourselves": The role of objects in national identity*. Leicester University Press, London, 1994.

[LEM 89] Lemaître, R., & Roux-Fouillet, P. Suzanne Briet (1894-1989), *Bulletin d'Informations de l'Association de Bibliothecaires Français*, 144, pp. 55-56, 1989.

[LEV 94] Levy, D. M. Fixed or fluid? Document stability and new media. In: *European Conference on Hypertext Technology 1994 Proceedings*. (Pp. 24-31). Association for Computing Machinery, New York, 1994.

[OTL 20] Otlet, P. *L'organisation internationale de la bibliographie et de la documentation*. (IIB Publ. 128). Institut International de Bibliographie, Brussels, 1920. Translation in [OTL 90: pp. 173-203].

[OTL 34] Otlet, P. *Traité de documentation*. Editiones Mundaneum, Brussels, 1934. Reprinted 1989, Liège: Centre de Lecture Publique de la Communauté Française.

[OTL 90] Otlet, P. *International organization and dissemination of knowledge: Selected essays*. (FID 684). Elsevier, Amsterdam, 1990.

[PEA 90] Pearce, S. M., ed. *Objects of knowledge*. (New research in museum studies, 1). Athlone Press, London, 1990.

[RAN 63] Ranganathan, S. R., ed. *Documentation and its facets*, Asia Publishing House, London, 1963.

[SCH 35] Schürmeyer, W. Aufgaben und Methoden der Dokumentation, *Zentralblatt für Bibliothekswesen*, 52, pp. 533-543, 1935. Repr. in [FRA 78: pp. 385-397].

[SEB 94] Sebeok, T. A., ed. *Encyclopedic dictionary of semiotics*. 2nd ed., Mouton de Gruyter, Berlin, 1994.

[SHE 72] Shera, J. H. *The foundations of education for librarianship*, Becker and Hayes, New York, 1972.

[SHO 77] Shores, L. *The generic book: What is it and how it works*, Library-College Associates, Norman, Oklahoma,1977.

[VOO 64] Voorhoeve, N. A. J. F. Donker Duyvis and standardization. In: *F. Donker Duyvis: His life and work*, (NIDER publ. ser. 2, no. 45). Netherlands Institute for Documentation and Filing, The Hague, pp. 39-50, 1964.

[WAR 90] Warner, J. Semiotics, information science, documents and computers. *Journal of Documentation*, 46, pp. 16-32, 1990.