How to Teach About Information as Related to Documentation?

by Henning Spang-Hanssen

The Concept of Information: A Preface to the Article by Henning Spang-Hanssen

In 1970, Professor Henning Spang-Hanssen held a speech entitled How to teach about information as related to documentation? I got a copy of the manuscript at about the same time because Spang-Hanssen served as mentor on my master’s thesis on scientific communication. This manuscript has not been published until now. I have kept it for all those years, often used it and referred to it (it is therefore also cited in the Social Sciences Citation Index). It has influenced my own thinking about information, and I still regard it as a valuable contribution.

Spang-Hanssen referred to this article as belonging to "the documentation era". Why do I consider this old paper important today? First of all, there is still much confusion about the concept of information and its place in library science and documentation. This confusion concerns whether information refers to information technology (IT) as a way of transmitting knowledge or whether it means the thing being communicated, the content of the transmission process. There are important differences between using the concept of information or the concept of documents (whether electronic or not). A document has a history, an author, a history of influence and reception in other documents, while "information" tends to be ahistorical and unsituated. "Information" is an important concept to help raise the status of a dusty library profession, as Spang-Hanssen writes, but it is problematic as a fundamental concept in LIS.

As a theoretical concept, "information" tends to move LIS unto theories about control, feedback, coding and noise in transmitting messages, while "document" tends to move LIS towards theories about meaning, language, knowledge, epistemology, and sociology. In LIS there may therefore be a whole paradigmatic conflict hidden in those words.

What I like most in this paper is this passage:
Information about some physical property of a material is actually incomplete without information about the precision of the data and about the conditions under which these data were obtained. Moreover, various investigations of a property have often led to different results that cannot be compared and evaluated apart from information about their background. An empirical fact always has a history and a perhaps not too certain future. This history and future can be known only through information from particular documents, i.e. by document retrieval.

The so-called fact retrieval centers seem to me to be just information centers that keep their information sources – i.e. their documents – exclusively to themselves.

A final argument for publishing this paper in Human IT is that Spang-Hanssen’s conclusion very clearly points to the need of basing LIS as a HUMAN science.

Birger Hjørland

About the Author

Selected Bibliography [works by Henning Spang-Hanssen]

The word information is certainly not a new word in the English language – nor in other languages that adopt Latin words, for instance German (Information) or Russian (informacija). However, during the last 20 years this word has acquired the status of a word very much in vogue – like for instance communication, data and cybernetics.

So in Denmark the word information was certainly known by at least certain groups of people before the Second World War, but nowadays it has almost suppressed the formerly much more used Danish word "oplysning" (which etymologically corresponds to enlightenment). This development is due to a variety of influences of a cultural, political and technical nature, and it has been a result of the word being used extensively in professional terminology and in the language usage in newspapers and radio.

The word information has been successful not only as a separate word, but even in combinations like information theory, information system, information center, information retrieval, and SDI (selective dissemination of information). The term information science is rapidly gaining ground in the United States as covering what has earlier been spoken about as documentation research or theory of documentation. It is illustrative that the American Documentation Institute changed its name a few years ago to the American Society for Information Science.

Correspondingly, in Europe the term informatik (German Informatik, French informatique, Russian informatika) is gaining ground in the same sense, together with for instance German Informatiker to replace Dokumentalist. In practical respects the expression Informatik is definitely preferable to Dokumentationswissenschaft, but unfortunately the terms Informatik and information science may be misleading as to
the sense of the terms. Information in the sense relevant to our field – whether this field is called Documentation or Information Science – seems to me definitely connected with documents, or more specifically: with the content of documents, i.e. what is said in documents. It goes without saying that documents are here meant to include not only texts, but even recorded speech, and drawings, diagrams etc. in accordance with common terminological usage in our field.

The terms *documentation* and *documentalist* point by themselves to this connection with documents, whereas the terms *Informatik(er)* and *information science* give rise to confusion with terms like *information theory*.

*Information theory* is an unfortunate – but since 20 years well established – designation for the statistical theory of communication developed in the teleengineering field by Nyquist, Shannon a.o. This field is not concerned with documents, and not even primarily concerned with the content or meaning of documents or other symbolic representations, but concentrates on the efficient transmission of signals, which may – or may not – convey meaning. It is therefore unfortunate to confuse the term *information theory* with information as occurring in information science and information retrieval.

Moreover, these terms are not seldom confused with a more or less obscure use of the word *information* to mean something *factual* or *real* as opposed to representations of such facts; what is found written in documents – or what is said in a lecture – are according to this view only disguises or at best surrogates of facts. This more or less vague conception seems to be the basis of the distinction sometimes made between "fact retrieval" and "document retrieval".

This distinction I find philosophically unbased; we here touch upon the fundamental problem of the meaning of meaning and of the nature of signs and symbols. What is more essential to us, this distinction seems unfortunate in actual documentation work. There will, admittedly, be cases in which a document or information center is set up with the exclusive function of providing information concerning physical data, or statistical figures, or exchange rates of currencies, or stock market prices. But even in such cases, it applies that neither the person who requests such information nor the person who delivers it should ignore the reliability of data and forget about the general setting in which the data is acquired. Information about some physical property of a material is actually incomplete without information about the precision of the data and about the conditions under which these data were obtained. Moreover, various investigations of a property have often led to different results that cannot be compared and evaluated apart from information about their background. An empirical fact always has a history and a perhaps not too certain future. This history and future can be known only through information from particular documents, i.e. by document retrieval.

The so-called fact retrieval centers seem to me to be just information centers that keep their information sources – i.e. their documents – exclusively to themselves.

On the other hand there has been a tendency during the last decade to set up information centers and services that act only as centers for information *about* documents, not as centers for information *from* documents. In extreme cases, such centers only provide the requester with a list of references to documents supposed to
deal with the subject matter he is interested in; no care is taken to provide him with the actual documents, and perhaps not even to provide him with information about where to get hold of the documents listed. Paradoxically enough such document retrieval centers or services may not be in possession of documents at all, except for secondary documents, i.e. bibliographies, indexes and abstract publications. Correspondingly, the term documentation is sometimes taken to mean the art of providing references, or at best: providing surrogates like abstracts or extracts.

In the Annual Review of Information Science and Technology (1969), Herbert B. Landau deals with this situation in the chapter on document dissemination (esp. pp. 234–235). As pointed out there, this tendency to identify documentation with the providing of references, and possibly abstracts, has been reinforced by the increasing use of automatic data processing (ADP). Many people in various fields have pointed out that calling in computers means solving exactly those problems that computers can solve – at the expense of the remaining, perhaps more fundamental problems. ADP services are clever at providing references; well then – let us identify providing references with documentation work or information work. Some people even identify information science with the study of automatic data processing in documentation and library work.

The two tendencies I have discussed should be regarded together: when some people identify documentation with providing references – not documents, and definitely not the content of documents – it is no wonder that some other people look for information retrieval of another kind – concentrating at the extreme on the factual content of documents – and no wonder a particular term has been introduced for this reaction, namely "fact retrieval".

Both extremes may have their proper fields: a so-called fact retrieval system may be useful, e.g. as an internal member of a large industrial firm, or of a large administrative institution, while a reference providing system may be useful, e.g. as an auxiliary service common to cooperating research libraries. In all such cases the immediate users are professional and highly qualified people, who know what the services are good for. But in the more general cases of information centers or documentation centers that serve a variety of research workers, of industrial needs, of educational institutions and the public in general, certain consequences can, however, be drawn from the discussion above:

Firstly, when the content of documents is regarded as essential to and characteristic of documentation or information work, the core information centers must be institutions very much like research libraries as we know them, e.g. here in Denmark. Modern information work seems to me an updating of special librarianship, but it shall be stressed that this updating is not obtained without constant effort.

Secondly, it is unsatisfactory to provide the users only with so-called facts, or only with references (which actually are facts also, namely of a bibliographical nature). We are to inform the user about the background of these so-called facts, and we are to assist him in evaluating and utilizing the documents upon which the facts rely. But in order to do so we must know the users, i.e. we need to know more about the human factor in information.
Scientists and engineers can be provided with the finest of information services, but they will not necessarily use these services, nor will they engage in the painful effort of reading and, even worse, thinking about the materials with which such services can provide them. The indexing and retrieval of documents present far fewer problems to the designer of information systems than does the study, analysis, and modification of the human behavior patterns of his clients. (Ladendorf 1970, p. 215)

The renewed interest in the part played by human nature is also the subject of a stimulating paper by M.B. Line (Bath Univ. Of Technology) in ASLIB Proceedings (1970). Mr. Line points to various advantages of information systems that have not been too formalized (cf. e.g. p. 329).

We shall definitely not think of informal, human-oriented information systems as an easier goal to reach than highly formalized and apparently smoothly acting systems and services. On the contrary, to develop and maintain an efficient human-oriented system is a job that calls for much more research and much more daily supervision than to run formal, preferably mechanized systems. In fact, the essence of Mr. Line’s paper is to stress the necessity of extensive studies of users’ needs and habits.

I think that the most important point to make when one has to teach about information in relation to library and documentation work is to stress the part played by human beings in all stages of information work. Or to put it somewhat differently: if we try to set up a definition of information – or at least try to characterize what is meant by information – we will not succeed if we do not include human elements like user, author, ourselves as information workers, and the human language we speak and write.

When one is to teach about information, it may in fact be tempting to start with some formal definition. You will note that in this discussion I have resisted this temptation. Of course, our students and listeners have a right to know what we are teaching about, but I will recommend the tactics of encirclement before discussing some formal definition of information. The difficulty is not that definitions are not available; on the contrary it is precisely the variety of proposed definitions that should warn us from picking out one formal definition – or worse still: from adding to the variety by setting up a formal definition of our own. We should be careful before adopting the word information as a technical, i.e. professional term, with fixed semantic relations to professional terms like document, classification, indexing, thesaurus, abstracts, cataloguing etc.

In fact, we are not obliged to accept the word information as a professional term at all. It might be that this word is most useful when left without any formal definition, like e.g. the word discussion, or the word difficulty, or the word literature. It might be that the word information is useful in particular when we try to raise our professional status in relation to other professions; it sounds smart and imposing and gives an air of technicality. I find no moral objections to this sort of use of words; language is certainly not only for informative uses ("informative" here refers to the so-called intellectual or factual meaning of a text or utterance). However, we must realize that the status-increasing effect of a word may depend precisely on its being used in other fields as well, preferably in fields that have a high status, like engineering and, nowadays,
sociology. The uses in such other fields actually makes it impossible to at the same time keep this word as a formally defined professional term in our field without some risk of confusion; the words *force*, *energy* and *effect* – used both generally and in physics as formally defined terms – illustrate this situation.

The word information – and combinations like information retrieval, information center – have definitely contributed to raise the public opinion of library and documentation work, which is generally held to be a little dull, dusty and distant from what is actually going on in society. Maybe it should be wise to leave the word information there, were it not for the fact – already mentioned – that several attempts have been made to define information as a formal term relative to documentation and information work, and there have even been attempts to define information as some measurable quantity, corresponding to questions of the type: *How much* information was retrieved by the search?

Therefore, some discussion of the more or less formal uses of the word information is indispensable when teaching about documentation; the first point to make may be that the more formal the discussions and the definitions found in the professional literature, the more the authors in question stress that information is a relative quantity:

> It is very important to note that information is a relative quantity and cannot be defined except in terms of a specific situation with a specific set of observable actions. Then, and only then, can we define a quantitative measure for information. (Yovits 1969, p. 19)²

This relativity of information can be illustrated and further examined by pointing to how the word is used in ordinary language, here English.

Primarily for reasons found in the historic development of European philosophy, professional definitions are usually dealing with nouns (substantives), rarely with other parts of speech. We are accustomed to define e.g. *rain* by analysing a phenomenon, a concept, an idea or even a thing called "rain", not by regarding some act or event. At school we are asked "What is rain?", not "Define rain by describing what is going on when it is raining", and if a schoolboy answers to "What is rain?" by saying "Rain – that is when you get wet" or "Rain – that is when you need an umbrella", the teacher may scold him for being foolish and ignorant. This even applies to nouns that are derived from verbal stems, like *communication* – *communicate*, *information* – *inform*; the traditional way of giving definitions – at least in Indo-European speaking countries – here leads to questions of the type "What is information?" instead of "What do we mean by saying that someone is informing someone else about something?"³

In order to study the relativity of what is called information it is, however, a definitely more fruitful strategy to start from questions of the latter type. Starting in this way, we include from the outset

- A person acting as informant
- A person (possibly several) being informed (here called: informee)
- Something being informed about
And by adding to the question above – or just by following up the listing of possibly relevant factors or conditions – we include

- a possible result or outcome of the act of informing
- a possible failure of the act of informing
- a possible intention (of the informant) to inform
- a possible intention (of the informee) to be informed
- a possible intervention by someone or something. In the positive sense, e.g. in the case of a documentalist assisting a research worker in being informed by some document by pointing to it. In the negative sense, e.g. in the case of electronic or acoustic disturbances – noise – on a telephone line transmitting news.
- an oral or written message produced by the informant, possibly in some technically coded representation.

It seems possible to add more factors, and to specify some of the factors above into a number of sub-factors. Actually, while the word information taken in isolation might lead to too narrow a conception, we now seem to run a risk of getting involved in too complex a situation. Obviously, it is necessary to choose some factors as essential, if we are to define even vaguely a concept of information, and, obviously, the definitions arrived at will vary according to which, and how many, factors are chosen as the essential ones. So, in the so-called information theory or statistical theory of communication only the message and, more specifically, coded representations of messages – together with possible technical disturbances – are essential to the definition of information, or to be precise, of amount of information.

In the following, I shall discuss some conceptions of information in relation to documentation, in particular such uses of the word information that point to measurable quantities. It is in perfect accordance with the ideal of some natural sciences that we look for measurable quantities in order to find out to what degree we succeed in our efforts, in order to compare systems, and in order to describe the growth of literatures etc. No wonder the word information is often tacitly used in the sense "amount of information", and that – on the other hand – various measurable quantities are named "information", just because they are measurable (the quantity H introduced by Shannon⁴ is an example of this). Let us therefore discuss such quantities on the basis of a scheme that includes – at most – four of the above-mentioned factors, viz.

INFORMANT
INFORMEE
INFORMATION (about something)
DOCUMENT (cf. handout)

As long as the INFORMANT and/or the INFORMEE are taken into consideration, INFORMATION can be used to designate an act or process. If both (kinds of) persons are left out of consideration, INFORMATION is (usually) understood as
If *to inform* (or INFORMATION as an act or process) is to mean something other than to talk or to write, the INFORMEE will be an indispensable factor (even though he need not be known by name). In other words: a person cannot reasonably be called an INFORMANT unless he has at least *an intention* to INFORM (someone). Nevertheless, one of the most popular expressions relating to documentation work, namely the INFORMATION EXPLOSION, disregards the INFORMEE(S). What is called the information explosion can in the first place be termed only the *publication explosion*, or even the paper explosion: the number of printed pages in professional journals and books is increasing at a rate that can be described by an exponential function, like explosions. This, however, does not form an explosion of *information*, unless the number of printed pages is proportional to the amount of information resulting from the production and the distribution of these pages. In other words, when using the expression "the information explosion" we tacitly assume that professional papers contain information to a constant degree, regardless of their number, and regardless of their being utilized by informee(s).

The underlying conception of information is not particularly useful. It might be, e.g. that the users are able only to utilize a limited amount of literature, regardless of how much literature is produced; in that case the total outcome of information processes cannot exceed the limit set by the informees, and no information explosion can take place. One might even imagine, that an explosion-like growth of produced literature would have a lowering effect on the total utilization of the literature, i.e. would tend to decrease the total outcome of information processes: people could react as if they were being choked.

Even in other respects the growth of the number of printed pages seems to be too primitive a measure of information. Scientific and other professional papers are not produced exclusively for informative purposes, but also as tokens of activity and as means of increasing some status; the need to publish – in order not to perish – seems to play a more important part than earlier. It should also be noted that professional papers become obsolete as means of information at a much quicker rate than earlier. This means that even if the number of pages per year doubles every ten years, the total number of pages relevant today may not to a correspondingly high degree exceed the total number of pages relevant ten years ago. The quick "death" of much professional literature is not to deplore; actually many professional papers are nowadays meant to be only temporary means in research and instruction – it is just a pity that they are not printed with vanishing ink! As documentalists we shall remember that users are normally badly assisted by obsolete information. The relativity of information even applies to time.

I, therefore, am very sceptical of attempts at measuring growth of information by the growth of literature. There are interesting conclusions to be drawn from studies of e.g. the growth of scientific periodicals (cf. Price 1956, pp. 240–3), but they do not necessarily apply to information.

The literature produced can be studied with regard to the authors. *Lotka* has observed a very unequal distribution of scientific papers among authors, in that a large fraction of a collection of papers were written by a small group of scientists, while the majority of
scientists had contributed only one or a few papers each.\textsuperscript{5} Again this is relevant to questions of information only in so far as the amount of papers (or pages) is proportional to the information resulting from the papers. Now, from the study of sociological patterns of science and of research work in general it is known that authors form groups with regard to mutual citations and references; the Science Citation Index can be used to examine this in detail. It is plausible that even readers and hence users of papers form groups with regard to preferred authors, which would mean that very productive authors are not only quantitatively, but even qualitatively highly esteemed by certain groups of users. Such users will have a tendency to utilize information from this source, and the users’ habits are important to us in order to assist users. In this way, studies like Lotka’s, only more immediately focused on the utilization of papers, will tell us something about how to measure information.

It is interesting that a distribution very much like that observed by Lotka has been observed by Bradford in a different field, viz. regarding the distribution of papers or articles about some given subject matter on various periodicals.\textsuperscript{6} For a given subject or theme, like vitamins, or thermodynamics, the great majority of papers are published in a core of very few periodicals – different ones, of course, according to subject – while on the other hand one or a few papers per year can be found scattered in a variety of periodicals. This type of distribution, known as Bradford’s Law, is extremely important for the planning of acquisition of periodicals in libraries and information centers. It means, that small or highly specialized libraries should possess the core periodicals of the subjects relevant to the library whereas only National Libraries can afford to cover peripheral publications; in fact, complete covering is hardly worth aiming at in smaller countries, and the solution is worldwide co-operation.

Now, this is a result of great importance to documentation, but once more we must not identify articles, i.e. literature, with information unless we include the attitude of users and the actual usefulness of various articles to users. It is not immediately warranted to say that the information on a particular subject is scattered in the same way as the articles. We must specify the INFORMEE and his particular situation; probably he is not ignorant beforehand on the subject he is interested in, and if, on an average, the more peripheral periodicals will publish more specialized papers on the subject in question than the core periodicals, he will probably find some of the peripheral periodical papers more informative. But the reverse may also be the case. Once more, I am not aiming at diminishing the importance of statistics and the measurement of literature; on the contrary, I should like to encourage work of this kind, because we know too little about everything. But precisely because we know too little, we should not be content with rash generalizations or vogue words.

Bradford’s distribution applies to articles regardless of INFORMANT and INFORMEE, but with regard to the subject matter or theme of the document. Other proposed measures of information do not even take this into account, but actually deal with some internal structure of a document or message, regardless of meaning. This – as already mentioned – applies to the concept "amount of information" in Shannon’s sense, i.e. in the so-called information theory. The amount of information is here measured by the decrease of uncertainty resulting from the choice of a particular message among a set of possible messages. This sounds as though it has something to do with information in its vague and usual sense; cf. the choice of the message
Wednesday among the set of the seven names of the days of the week, in order to tell what day it is today. It would take us too far to describe the background of Shannon’s "amount of information". I shall only mention a few points to show the limitation of this measure to our conception of information.

- In Shannon’s sense, the amount of information is proportional to the length of the message (in a given code). This obviously does not apply to the utilization of literature as information. Among other things, an abstract may be as informative as the complete paper.

- Shannon’s amount of information presupposes a measure of the uncertainty on behalf of the receiver. By the utilization of literature as information no measurable uncertainty can be defined generally.

- Shannon’s amount of information applies to some explicit coding and cannot in the case of normal writing (or speech) account for semantic relations that are not shown by similarities of expression. E.g. the synonyms "serials" and "periodicals" would be treated as different messages (or parts of messages) having different "amounts of information".

From the last remark it will be seen, that Shannon’s "amount of information" may be applicable to the vocabulary of a thesaurus or other fixed and controlled vocabulary used for indexing and retrieval purposes; in fact, it seems useful and necessary to study thesauri from this point of view. But thesauri are not immediate sources of information, and once more I conclude, that the measure of information discussed has some bearing on information in our sense, but actually deals with other, more specialized or formal concepts.

(To the top)

About the Author

Henning Spang-Hanssen (born 1920) graduated as a M.Sc. in Engineering in 1943. He became Dr. of Philosophy in 1960 with the dissertation Probability and Structural Classification in Language Description. Between 1943–1965 he was employed in various industrial companies. From 1965–69 he was a research librarian at the National Technological Library in Denmark, and from 1969–1990 he was professor of applied and mathematical linguistics at the University of Copenhagen. In this position he started a new line of study concerning "Linguistic Aspects of Information and Documentation". In 2000 he became Dr. of honour at the Copenhagen Business School. He has been a member of the national council for humanities research, of DANDOK and NORDINFO and of national and international commissions related to problems of language and terminology. His is one of the very few people who have worked in a research library and tried to relate this to theoretical problems in documentation or information science. In 1999 and 2000 HSH served as a member of evaluation committees for the selection of qualified candidates for Professorships in Library and Information Science in both Norway and Denmark.
Selected Bibliography [works by Henning Spang-Hanssen]


Of special interest in this connection are the following works by Henning Spang-Hanssen:

1959: Probability and Structural Classification in Language Description (Dissertation).
1974: Roles and Links Compared with Grammatical Relations in Natural Languages. (Dansk Teknisk Litteraturselskab, nr. 40). Lyngby: Dansk Teknisk Litteraturselskab.


(To the top)

Notes

2. This article was also published in *American Documentation*, where the quotation can be found on p. 374.
3. [Cf. Machlup (1983, p. 657): "the use of the term information in both living and nonliving systems [is] acceptable as long as one does not forget that the term is used as a metaphor. Real information can come only from an informant. Information without an informant – without a person who tells something – is information in an only metaphoric sense". Spang-Hanssen here expresses the same view that Machlup independently expressed about 10 years later. Note added BH].
4. [Shannon & Weaver 1949 (Note added BH)]
5. [As first reported in Lotka 1926. (editor’s note)]
6. [In Bradford 1948. (editor’s note)]

(To the top)

References


(To the top)

© Henning Spang-Hanssen 2001

Åter till Human IT 1/2001