The Empirical Approach in the Evaluation of Information systems

Robert Fugmann[1]

Dedicated to Dr. Ingetraut Dahlberg
on the occasion of her retirement from ISKO presidency,
in recognition of her merits for the foundation of ISKO
and of her untiring efforts for its sustainment

Dr. Robert Fugmann, born in 1927, is a chemist and information scientist. He has been active in the development of the theory and practice of indexing and information supply for 45 years. He has served as ISKO Vice President. Visiting Professor and lecturer at various universities and other institutions in Germany and in the United States. He is author of 5 books, 109 articles, and 13 book reviews and is recipient of awards from FID, the Association of German Chemists, American Society for Information Science, and the American Chemical Society.

Fugmann, R. (1999). **The Empirical Approach in the Evaluation of Information systems.** *Knowledge Organization, 26(1).* 3-9. 17 refs.

ABSTRACT: The comparative evaluation of different mechanized information systems continues to constitute a controversial topic in the literature. Diametrically different opinions, seemingly corroborated through empirical evidence, have been presented since the time of the Cranfield experiments.

Similar situations have often been encountered in the history of science if reasoning has exclusively been based on empiricism. In the information scientific field, several "empirical laws" have been formulated, for example that of the alledgedly inverse precision-recall relationship in information retrieval, of the assumed direct relationship between consistency and quality of indexing, and of the alledged equivalence of automatic with intellectual indexing.

Empiricism is seen as only another variation of positivism, which has been abandoned in the natural sciences since the middle of this century for its evident inadequacy, but has latently survived in information science and even now dominates here. Here, it constitutes a source of continual confusion and an impediment to progress.

For literally anything an empirical "proof" can be submitted provided that suitable examples are selected and methods are chosen. Substantial advance in Library and Information Science requires abandoning empiricism. Budd's "hermeneutic phenomenology" seems to constitute a promising substitute (1995).

Table of contents

1. Introduction

The goal of information systems is to create order in a collection of documents so that an information searcher need not scan the entire collection in an attempt to find information of interest. As a consequence of the steadily increasing demands made on our information systems, they have been subjected to the necessity of being incessantly developed into higher and higher levels of advancement.

Development work must be oriented towards predetermined goals. The extent to which these goals have been approximated must be assessed through the application of quality criteria. If *wrong goals* are established and pursued the true goal can never be achieved. Should an information system's develop-

ment begin to stagnate or to depart from its goals, a new orientation is necessary. Otherwise, development work would yield increasingly unsatisfactory results.

The kind of philosophy that underlies development work plays an important role for the prospects of this kind of work. An inadequate philosophy may well lead research and development into the wrong direction and render it futile. Information system development does not constitute an exception to this rule. Positivism is an example of such an inadequate philosophy.

2. The dominance of the positivistic view in information science

In the positivistic view, reality of phenomena is only acknowledged if they are *observable.* This philosophy flourished until the second and third decades of this century. It was abandoned for its obvious inadequacy, at least in the natural sciences. Even the idea of atoms and molecules was rejected by prominent positivists because they could not be observed and realized empirically.

Positivism, especially in its variation of empiricism, claims that all knowledge exclusively originates from observations and from experience made by human beings (discussed, for example, by Hjorland 1997, p. 59; 1998, pp. 169, 176). Empiricism relies on the observable (and on its measurement) and neglects what cannot be observed or is excluded from observation. Empiricism would, for example, certainly have supported the medieval statement of the direct relation of weight to speed of freely falling bodies: the heavier the body, the faster it falls. Ample *empirical* evidence for this thesis could be compiled by comparing the free fall of various stones and pieces of paper. Galileo stated the opposite relationship, namely, the *independence* of falling speed from weight and could also present confirming observations in the case of massive falling bodies.

Some reasoning, in addition to mere observation, would have corroborated Galileo and refuted the opposite statement: if two stones are tied together (and the weight of the body is thus increased) there is no reason that they should fall faster than when they remain isolated. Which of them would have to fall faster under these circumstances and pull the other one and why should this be the case? Obviously, something is wrong with the empirical interpretation of experiments with light bodies. Today we know that air resistance is the factor that disturbed Galileo's experiments. In the vacuum, i.e. after the elimination of air resistance, the speed of falling bodies is independent of their weight.

Empiricism leads us astray when we select the wrong experimental conditions or restrict ourselves to a one-sided selection of examples. Using the wrong instruments is also a common source of fallacy, namely, using only those that merely happen to be available (instrumentalism, cf. also Budd 1995, p. 300). Ample *empirical* "evidence" could also be supplied for a postulated relationship between a purse's weight and the amount of money it contains if only those conditions are chosen which are conducive to the statement.

In the information field we encounter an instructive example of the adherence to the empiricist view which has led to treacherous results: purely empirically, an inverse relationship between precision and recall seems to prevail in information retrieval. According to this "empirical law", precision cannot be improved without simultaneously impairing recall and the reverse was also observed to prevail. For decades this statement has constituted an excuse for inadequate information systems and has thus paralyzed their improvement, although this type of relationship has often been put in doubt (for example by Soergel 1985, p. 122; Harter 1990, pp. 136, 145; Green 1992, p. 87; Fugmann 1994, p. 154; Svenonius 1995, p. 247).

What is neglected here is that information systems are observed which work perfectly both at 100 % precision and 100 % recall, as is the case, for example, for the majority of chemical molecular structure searches, for searches in the telephone directory, and for searches in other fact databases.

It is true that an inverse relation between precision and recall is often *observed* but only under circumstances in which there is no *predictability* in essence selection and in essence representation during indexing. Under these unfortunate circumstances, one has to find one's search statements by way of trial and error. This lack of predictability occurs

- if there is no index language in use for general concepts

and/or

- if an index language is used in an unreliable manner.

The postulated "law" of the inverse relation has been stated in the early Cranfield experiments and has persisted until today.

In the information field and in a purely empirical approach,

> literally *any* (intended) evaluation result for an information system can be produced and *any* opinion on the effectiveness of a type of information system can be corroborated

through the choice of or through the adherence to a specific constellation of experimental conditions and, in fact, in a manner which seems perfectly convincing and unsuspicious, just as is the case in the experimentation with falling bodies, if there is no critical contemplation of the experimental conditions.

Hence, *empirical* "evidence" cannot constitute *genuine* evidence. But empiricism can, by providing facts, give an incentive to a more advanced *interpretation* of reality, such that *observed* phenomena can be better explained and (yet) *unobserved* ones can be better predicted.

3 The inadequacy of consistency as a criterion of indexing quality

Traditionally, indexing consistency is recognized as a criterion of indexing quality, although this criterion has repeatedly been put in doubt (cf., for example, Cooper 1969; Soergel 1994, p. 594). This statement neglects the fact that a mode of *perfectly consistent* indexing may well be consistently defective. An optimum of indexing consistency could easily be achieved through the application of automatic indexing in any variation, for example through merely mechanically extracting keywords from natural language texts using a stop word list. But this may well lead to an ineffective or even unusable information system because it leads to unpredictable expressione for the concepts of interest.

On the other hand, and intuitively, consistency has *something* to do with indexing quality. If the same document is indexed entirely differently by different indexers (or indexed differently by the same indexer at different times), something seems to go wrong.

The solution to this puzzle is that consistency is paramount only in the *first step* of indexing, that of *essence selection*. Here it is identical with the *predictability* of the selection. In the *second step* of indexing, that of *essence representation*, only *predictability* is required (Fugmann 1993, p. 94-97).

A concept may well have been entered into the search file with different modes of expression, i.e. markedly inconsistently. But if these expressions can be looked up, they are made *predictable*. These expressions can then be compiled into a set of alternative search statements. For example, a substance may be represented by a variety of expressions, such as "vitamin C", "ascorbic acid", "cantan^{R}", or a paraphrasing expression like "scurvy-preventing substance in vegetables", etc. But if we know these expressions, we can readily retrieve the corresponding documents.

Hence, consistency is an invalid criterion for indexing quality. What is needed, instead, for good retrieval is predictability, both at the stages of essence selection and essence representation.

4 The number of access points as an invalid criterion for expected retrieval success

It has been postulated that the greater the number of "access points" to a document, the better its retrievability will be. If this is true, full text search files should be ideal for high recall values because they provide the maximum number of access points. But we know that full text files are far from being ideal in this respect (cf. for example, Blair 1996, p. 19). The reason is that the *words of the searcher* for the topic of interest only rarely match the *wording of the documents* of interest, especially in the case of searches for general concepts and topics.

Only a type of expression that is phrased in a *predictable* mode can constitute an access point for retrieval, a condition which is not fulfilled in ordinary natural language. Therefore, indexing languages, with their capability of providing *predictable* concept representations, may well be more specific *in retrieval* than an even more detailed natural language text, although an indexed text presents substantially fewer "access points" for the mechanized search, especially in the case of searches for *general* concepts and topics.

5 Survival power: the neglected criterion of information system quality

It is alien to the empiricist view to look into an information system's future because its future cannot be observed and because the system's fate cannot be empirically demonstrated in the present. What is generally preferred is a snapshot-like analysis of the *presently prevailing* situation because such an analysis is fast and cheap.

However, an information system user should be highly interested in whether an information system is *capable of continuing its service in the future*. There may be various reasons to abandon an information system in the future after some time of having been practiced in operational use. The demands made on retrieval precision may be steadily growing during the growth of a mechanized information system and the system may fail to provide significantly higher search specificity. Then it begins to produce hundreds or even thousands of irrelevant responses, from which the few relevant ones have to be sifted out through human inspection. This is a procedure which may well develop into requiring an intolerably high expenditure of time and attention.

An information system may also become unusable because its vocabulary has become chaotic in the course of time. This renders the indexing procedure correspondingly unreliable and the search results become increasingly incomplete as a consequence.

This decrease in recall may remain hidden for quite a long time because it is difficult to observe. But when this deficiency becomes apparent some day, perhaps through a correspondingly extended and careful investigation, an information system suddenly loses all the appreciation that it had so far enjoyed. Most often, the reason for this failure is a misplaced parsimoniousness in the input step. This type of positivistic and treacherous user satisfaction has often prevented a more advanced and durable information system from coming into existence or from surviving its initial stage.

The notion of information system *survival power*, although coined early in the history of our field (Harmon 1970), has not become popular in contemporary information science literature, although this is one of the most important quality criteria for operational information systems.

6 The full text information system

The equivalence of full text information systems with intellectually indexed systems has often been stated and empirically "proved", for example in the Cranfield experiments. In fact, an equivalence (or sometimes even a superiority) seems to prevail, *but only if any distinction between individual concepts and general concepts is dispensed with* and when, due to the smallness of the experimental files, the experimenters can memorize the modes of expression for general concepts that happen to be in the file. Under such circumstances a crucial quality criterion, namely, the *predictability* of the mode of expression of a concept of interest, is not put to the test.

Here we encounter what has been appropriately called the "small system syndrome", that is the phenomenon that *small* information systems display properties that are *fundamentally different* from those of large ones ("non-scalability" of information systems) and that the mere growth of a small system may well lead to its decline and eventual break-down (cf. Gey, Dabney 1990).

What is hidden to the empiricist view and what is therefore neglected in many evaluation studies is the fact that a *word* used in natural language is not intended for use in isolation. It is only in context that a word assumes meaning and importance and, to the reader or to the listener, the context is presented. Hence, in colloquial discourse, any text requires (and subconsciously receives) *interpretation* (cf., for example, Budd 1995, pp. 307, 308).

Text interpretation yields

- reliable essence recognition (the importance of a concept depends on the context in which it is embedded),
- word meaning disambiguation (through the knowledgable utilization of the word's context),
- paraphrase lexicalization (i.e. substituting a *paraphrase* in storage or retrieval by one of its *lexical* equivalents in natural or artificial language, such as notations or descriptors. The problem of expression multiplicity for a concept in natural language is much more than merely one of synonymy),
- ellipses filling (i.e. making explicit what has only been implied in a text and must be inferred from it, (cf. for example, Ranganathan 1962, p.129; Green 1992, p.84; Fugmann 1993, pp. 64, 70, 91),
- establishing concept relationships,
- (near) synonym control,
- verbalization of non-textual information,

all of which are typical achievements of good intellectual indexing. Thus, concept representations are made predictable, and concepts are made easily retrievable. However, these achievements are renounced in searching non-interpreted full texts, much to the detriment of search quality.

As far as *context* is concerned, it is always freely phrased and, hence, expressed in an *unpredictable* manner. Therefore, it escapes inclusion in the query as a reliable, interpretative statement.

Text interpretation defies mechanization because it is an inherently *indeterminate process*. Its point of departure is an indeterminate one, namely, the unpredictable mode of expression encountered in natural language. Hence, interpretation proceeds in an unpredictable manner, too. No instructions can therefore be laid down *in advance* in a program for the satisfactory, mechanized execution of interpretation. These instructions would have to be infinitely numerous.

Full text storage has often been given credit for the specificity which it seems to provide for searches. Here it seems (and occasionally is) superior to indexing and classification, which often lack sufficient specificity. But

> *specificity without predictability is largely useless for searching*

as is obvious from the foregoing. Using this specificity will yield both false responses (through the unresolved ambiguity of natural language words) and a dramatic loss of information.

7 User evaluation

Any indexing must aim at satisfying its users, not only in the present, but particularly in the future and also under the changing requirements of the future. It is therefore paramount for an information system that the users are involved in the design of the system

and in continually adapting the system to unforeseen new requirements.

Specifically, users must render their opinion as to whether the conceptual categories of the system under discussion meet the requirements for searches and whether navigation in the vocabulary is sufficiently easy and fast. Only in passing we mention here that users are often unaware of the necessity of navigating in the vocabulary. They often omit this opportunity of phrasing the set of search statements as completly as possible and, thus, the opportunity of making the set of responses from the search file as complete as possible, too.

Users must give feedback on whether the specificity of the vocabulary should be extended or reduced. They should urge for the timeliness of the input and for an appropriate speed in the execution of the searches. The searches should be affordable, and the texts of the responses should be easily accessible. Users should also care about the maintenance of an appropriate coverage of their literature in their information system. They should express their opinion about the quality of book indexes to publishers in order to make them aware of the necessity of improving them. Their opinion in these affairs serves for the improvement of information supply from these sources.

In rendering their opinion on an information system users must also concede

- that not all present-day requirements could have been anticipated in the past and could therefore not have been taken into consideration through a correspondingly appropriate, provident selection of documents and indexing routines;
- that an information system must display some features the judgement of which goes beyond their competence, in so far as the user is not an information expert. (Some of these features have been discussed in the foregoing.)

In particular, users should express their opinion on the information system's performance with respect to precision and recall. Here, a widespread misconception in information system evaluations is caused through the neglect of the difference between a *delegated* and a *non-delegated* search and through the confusion of *relevance* with *pertinence*, a difference which has early been emphasized, for example by Kemp (1974).

In assessing the retrieval quality of their information system, users must distinguish the (desired) *relevant* responses (i.e. those responses that *meet the elements of the search request*) from the undesired ones which do not match these statements. Among these there may well be *pertinent reponses*, i.e. those that happen to be *of momentary interest to an individual* (!) user but which *do not match* the search request.

What is *pertinent* for an *individual user* may well not be pertinent for the vast majority of the *other users*, and may even be pertinent for an individual user *only at a particular point in time.* But the user would be unable to specify *in advance* which (still unknown) records would be rated "interesting" or "non-interesting" when and if they are encountered.

When an information seeker searches the literature, he or she is accustomed to encountering a great variety of documents, many of which prove to be of *interest* although they are more or less *distant from the goal* that the searcher initially had had in mind.

It is inherent in one's *non-delegated* search that one has an entirely free hand to gather what raises one's interest, i.e. what is pertinent for a user in this sense, *irrespective of any predetermined search goal.* In the delegated search, on the other hand, *somebody else* decides what should be selected or rejected. The decison is made by the computer programer and/or the intermediary who assists in executing the search and in scrutinizing the responses.

Many users expect that a computerized search will "at least" yield the results of their non-delegated search. But any delegated search requires the *definition* of what is to be searched. Only what can defined can be satisfactorily delegated to another person or to a mechanism.[2] The computerized search does not constitute an exception to this rule.

Trying to *approximate* the pertinence goal may well lead an information system directly into its decline and eventual break-down in the more or less distant future (cf., for example, ISKO/IE 1992). Too many false, unrequested responses will soon be retrieved and bury the few requested and retrieved ones.

Satisfying the pertinence criterion for the responses to a computerized search would require the omniscient, clairvoyant, and perfectly prognostic programer and indexer. S/he would have to lay down *in advance* the instructions according to which the pertinence decision is to be made and what should therefore be submitted as a response to the user's query or should providently be included into the search file.

Hence, the pertinence expectation is inherently unsatisfiable in a system for the delegated information search because this type of search is typically oriented to what has been *laid down in advance* as the goal of the search.

What is attainable in mechanized retrieval and should constitute its genuine goal is the *relevance* of the responses (often also called "topical relevance") in the aforementioned sense, i.e. the selection of those documents *that satisfy the search statements* of the in-

formation seeker, however they may have been phrased, and *to exclude what had not been requested* by the user.

This holds true independent of whether the search is executed online and interactively or in batch, because the interactive search is merely a fast succession of searches each of which precisely follows the *mechanism* of the traditional batch search, i.e. the mechanism of mechanically scrutizing whether an item in the search file more or less completely matches the search statements of a query.

Hence, subjective, serendipitous browsing the literature continues to constitute the appropriate way of finding what proves to be pertinent, and in fact in a individual user's specific, idiosyncratic, and time-dependent situation. The computerized search cannot constitute a substitute for serendipitous browsing of the literature.

It is a misconception of "user friendliness" if the goal is to try to simulate the highly subjective and undefinable process of serendipitous browsing executed by the user. User friendliness should not be permitted to lead to the eventual break-down of the information system, much to the disadvantage of the same users who have initially relied on it and have insisted on such a type of "friendliness".

Here, the information specialist must be firm in contradicting such an ill-considered expectation on the part of the naive user and in contradicting the corresponding criticism of information systems that do not offer the full serendipity of the non-delegated search.

In assessing recall and relevance of an information system again a markedly positivistic-empiricistic attitude widely leaks through. Unretrieved, though relevant, documents are not (easily) observable and are therefore often neglected. Some authors even advise against worrying about information loss because it is hidden to the user and, hence, does not lead to dissatisfaction. Ironically, the less careful the assessment of an information system and the fewer the omissions traced here, the "better" the system comes out in such a purely empirical evaluation (see the criticism of Blair 1996, pp. 8-10).

Often, instead of assessing true recall values "relative recall" is used because it is fast and easy to observe and to assess. Here, the recall values of several information systems are compared. But it is no relief to a user, when some day all of these systems are unmasked to have been working at a miserable incompleteness of their responses.

Restricting evaluation to what can easily be observed is also a source of a wide-spread misevaluation of the *costs* of operating an information system. The figures for the input into a system or for its subscription are easy to assess and to criticize because they appear in the budgets. However, a (misplaced) input parsimoniousness causes high and continuously increasing (but largely hidden) costs in the use of the system, in particular through the expenditure incurred through sifting out the continuously increasing quantities of trash in the responses.

What is hidden, too, are the costs incurred through the omission of valuable but not retrieved information or through the loss of those responses that had escaped the searcher because they had been buried under unmanagable quantities of irrelevant responses.

Here, the wide-spread ill-reflected, purely empirical cost-benefit assessment has led to the proliferation of input-parsimonious information systems which, however, are increasingly losing their operability because of the steadily increasing (though latent) costs of using them. The situation here is aggravated through the fact that an information system's benefit is largely hidden, cannot be expressed in figures, and again escapes the appreciation of the empiricist and of the accountant.

It is true, information system evaluation must be based on empirical grounds, but this basis must always be carefully interpreted and reflected upon. In particular, the observations must not be of the snapshot type but must be continued over a sufficiently long period of time. It is only in the course of time, that an information system reveals its crucial features to the purely empirical view. Extending information systems' evaluation over time is among the recommendations in the review by Harter and Hert (1997, p. 63). The user is overtaxed if he or she is expected to undertake such expensive and extensive studies. The user will hardly feel responsible for an information system's failure which comes to light only after several years of operation.

Neural networks have been given credit for allegedly obviating any *a priori* instructions because they can autonomously "learn". But they can at best learn from *existing experience*. This is not sufficient for satisfactory processing what is continuously newly coming up in the literature and in the world.

8 Conclusion

A markedly positivistic - empiricist philosphy still dominates the field of information system evaluation and, hence, the design of information systems, as has emphatically been criticized by Budd (1995) and Hjørland (1997). This philosophy has led to the neglect of those important system features that are difficult to observe and to an emphasis on what can be easily observed (and even measured). The persistence of such a view throughout decades of research and

development is not surprising, as we know from the history of the sciences throughout the centuries. This, however, should not discourage researchers from defining new paradigms. Much more appropriate to the field of information science is the approach of "hermeneutic phenomenology" suggested by Budd (1995). Such a paradigm change promises to overcome several of the obstacles that presently stand in the way of progress in our field and may well constitute an incentive to substantial advancement.

Notes

1 Modified version of a paper presented at the 4th Congress of the Spanish ISKO section, Granada, 22-24 April 1999.

2 According to the axiom of definability: "The compilation of information relevant to a topic can be delegated only to the extent to which an inquirer can define the topic in terms of concepts and concept relations" (Fugmann 1985, p. 118 and1993, pp. 41, 45).

References

Blair, David (1996): STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years After. *Journal of the American Society for Information Science* 47, No. 1, 4-22.

Budd, John M. (1995): An Epistemological Foundation for Library and Information Science. *Library Quarterly* 65, 295-318.

Cooper, William S. (1969): Is Interindexer consistency a Hobgoblin? *American Documentation*, 268-278.

Fugmann, Robert (1985): The Five-Axiom-Theory of Indexing and Information Supply. *Journal of the American Society for Information Science* 36, 116-129.

Fugmann, Robert (1993: Subject Analysis and Indexing – Theoretical Foundation and Practical Advice. *INDEKS Verlag* Frankfurt/M.-Germany (now Ergon Verlag, Grombuehlstrasse 7, D 97 080 Wuerzburg-Germany). ISBN 3-88672-500-6.

Fugmann, Robert (1994): Galileo and the Inverse Precision-Recall Relationship: Medieval Attitudes in Modern Information Science. *Knowledge Organization* 21, No. 3, 153-154.

Gey, Frederic; Dabney, Daniel P. (1990): Letter to the Editor. *Journal of the American Society for Information Science* 41, 613.

Green, Rebecca (1992): The expression of syntagmatic relationships in indexing: Are frame-based index languages the answer? *Classification Research for Knowledge Representation and Organization.* Proceedings of the 5th International Study Conference on Classification Research, Toronto, Canada, June 24-28, 1991, ISBN 0 444 89 343 1.

Harmon, Glynn (1970): Information Need Transformation During Inquiry: A Reinterpretation of User Relevance. *Proceedings of the American Society for Information Science* 1970, Vol. 7, 41-43.

Harter, Stephen P.; Hert, Carol (1997): Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. In: Williams, Martha (editor): *Annual Review of Information Science and Technology* 32, 1-94; Learned Information Inc., Medford NJ/USA; ISBN 1-57387-047-1.

Hjørland, Birger (1997): *Information Seeking and Subject Representation.* Greenwood Press, ISBN 0-313-29893-9.

Hjørland, Birger (1998): The Classification of Philosophy: A Case Study in the Classification of a Knowledge Field. *Knowledge Organization* 25, 162-201.

ISKO/IE (1992): User Evaluation of Information Systems. *International Classification* 19, 151-152.

Kemp, D.A. (1974): Relevance, Pertinence and the Information System Development. *Information Storage and Retrieval* 10, 37-47.

Ranganathan, S.R. (1962): *Elements of Library Classification.* ASIA Publishing House, London. Soergel, Dagobert (1985): *Organizing Information – Principles of Data Base and Retrieval Systems.* Academic Press Inc., Harcourt Brace Jovanovich Publishers, New York. ISBN 0-12-654261-9.

Soergel, Dagobert (1994): Indexing and Retrieval Performance: The Logical Evidence. *Journal of the American Society for Information Science* 45 (8) 589-599.

Svenonius, Elaine (1995): Precoordination or not? *Subject Indexing: Principles and Practices in the 90's.* Proceedings of the IFLA Satellite Meeting held in Lisbon, Portugal, 17-18 August 1993. ISBN 3-598-11251-3.

Dr. Robert Fugmann, International Society for Knowledge Organization,
Woogstraße 36A, D - 60431 Frankfurt/M. (Office).

Private: Alte Poststrasse 13, D - 65510 Idstein,
Tel.: +49 6126 8250, Fax: +49 6126 4366