Is classification necessary after Google?[*]

Birger Hjørland

Royal School of Library and Information Science
6 Birketinget, DK-2300 Copenhagen S, Denmark
bh@iva.dk

Abstract

This presentation considers the nature of classification and the challenges that bibliographical faces in the digital age both in library practice and in information retrieval theory. It considers fundamental questions such as "how do establish that A is a kind of X?" and "how do we distinguish good from bad classifications?" The new trend of evidence based practice (EBP) is seen as an argument why classification is still necessary for scientific documentation, and the basic epistemological positions (empiricism, rationalism, historicism and pragmatism) is briefly considered in relation to classification.

Keywords: Classification, Bibliographical classification, Evidence Based Practice, Epistemology.

[*] This speech is based on a manuscript with the same title which has been submitted to *Journal of Documentation*.

1. WHAT IS CLASSIFICATION?

Classification is the interdependent processes of:

- Defining classes;
- Determining relationships between classes (such as hierarchical relations, among others), i.e., making a classification system;
- Assigning elements (in Library and Information Science (LIS): documents) to one or more classes in a given classification system.

This understanding of classification is fully in agreement with a definition suggested by Lois Mai Chan:

> [Classification is] the multistage process of deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or objects that have the property or characteristic in common into a class. Other essential aspects of classification are establishing relationships among classes and making distinctions within classes to arrive at subclasses and finer divisions (Chan, 1994, p. 259; also adopted by Golub, 2011, pp. 231-232).

This is equivalent to the interdependent processes of:

- Defining concepts
- Determining semantic relations between concepts
- Determining which elements fall under a given concept (to assign a given "thing" to one or more concepts).

This close relationship between classification and concepts indicates a close connection between classification research and concept theory (cf., Hjørland, 2009).

Basically classification is thus to say about something that it belongs to a given class or category and to say how that class is related to other classes. (Or rather - to anticipate my epistemological conclusion: Things do not "belong" to classes: They are *assigned* to classes by somebody from given perspectives and for given purposes. And similarly: classes are assigned relations by somebody from given perspectives and for given purposes and subjects are assigned to documents from given perspectives and for given purposes).

When we are speaking about classification of books, articles, pictures, music and any other kinds of documents, we are speaking about *bibliographic classification* (as opposed to the classification of "things" or "phenomena"). Bibliographic classification is mostly dependent on (derivative) of the classifications of "things" and phenomena. We normally in LIS classify books about plants, animals, countries, drugs etc. the same way as scientists or the broader community classifies plants, animals, countries, drugs etc.

Library scientist Henry Bliss (1935, p. 2) suggested that we should classify documents according *to scientific consensus* about classifying phenomena, but A. Broadfield wrote:

> Consensus is most likely to appear among the unenlightened, of whom it is characteristic to be unanimous on the truth of what is false. In intellectual matters agreement is rare, especially in live issues (Broadfield, 1946, pp. 69-70).

Whether we in LIS choses to follow Bliss or to agree with Broadfield, we need to describe the methods by which we examine the literature and draw our conclusions. In the first case: how do examine what the consensus is in a certain domain? In the second case: How do we make decisions in cases of lack of consensus? (All too often have both these questions been neglected, and just a practical, arbitrary decision has been made. Because of this the status of our traditional classification systems may be questioned. This may have worked well enough in the age of printed catalogues, but will not work in the age of Google).

In LIS "classification" is often opposed to "verbal indexing systems" as shown on Figure 1.

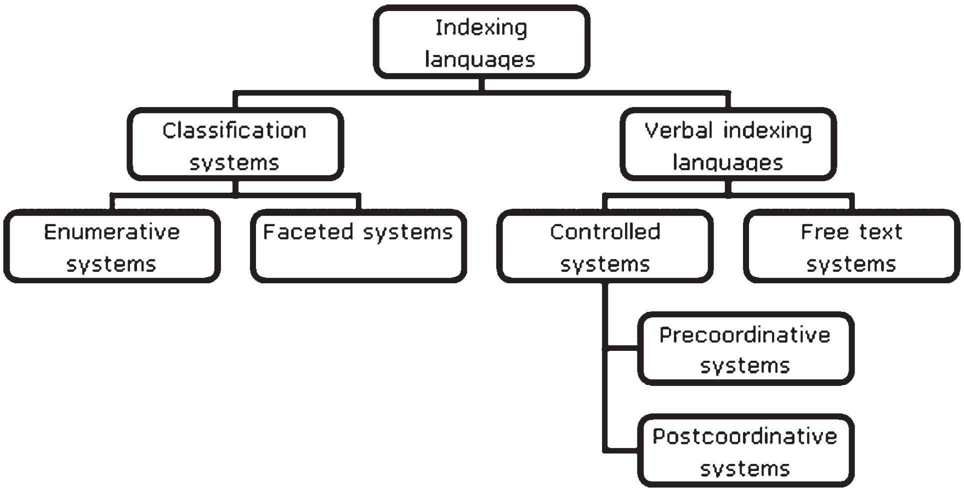

Figure 1. A traditional classification of indexing languages

The fundamental distinction in Figure 1 is between classification systems and verbal indexing systems. This distinction has been criticized by Lancaster (2003, pp. 20-22), who argued that "one should not speak of assigning classification codes as 'classification' in opposition to the assignment of indexing terms as 'indexing'." "These terminological distinctions –he writes–, are quite meaningless and only serve to cause confusion" (Ibid., p. 21).

The view that this distinction is purely superficial is also supported by the fact that a classification system may be transformed into a thesaurus and vice versa (cf., Aitchison, 1986, 2004; Broughton, 2008; Riesthuis; Bliedung, 1991). It is therefore important to realize that all of the different kinds of systems in Figure 1 - with the possible exception of free text systems - are different kinds of "classification systems."

An alternative to figure 1 is given in figure 2:

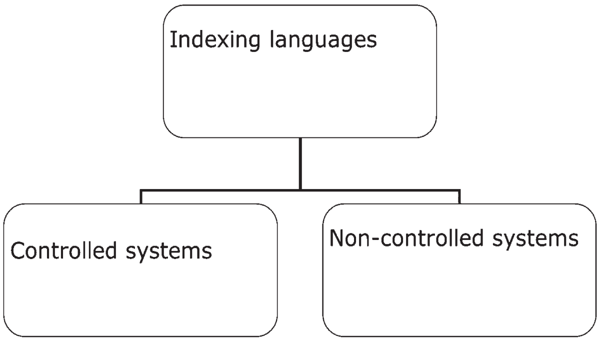

Figure 2. The basic classification of indexing languages

Hodge (2000) suggested a taxonomy of *Knowledge Organizing Systems* (KOS) which may also all be considered different kinds of classification systems:

Term lists
- Authority files
- Glossaries
- Dictionaries
- Gazetteers

Classifications and categories
- Subject headings
- Classification schemes
- Taxonomies
- Categorization schemes

Relationship lists
- Thesauri
- Semantic networks
- Ontologies

I shall not here discuss Hodge's taxonomy in detail. It should be considered, however, that because a faceted classification schema can be transferred to a thesaurus, it seems *not* to be based on fundamental criteria. The most important difference between the different kinds of KOS seems to be the different kinds of semantic relations being displayed. In traditional classification systems, hierarchical relations and the relationship between synonyms and homonyms are the most important. In ontologies, a large range of semantic relations are possible.

Some researchers have suggested that ontologies may be considered a form of KOS, which allow us to consider other forms of KOS as specific kinds of ontologies. Lars Marius Garshol wrote:

> With ontologies the creator of the subject description language is allowed to define the language at will… thesauri extend taxonomies, by adding more built-in relationships and properties. …
> A consequence of this is that topic maps [an ontology based technology] can actually represent taxonomies, thesauri, faceted classification, synonym rings, and authority files, simply by using the fixed vocabularies of these classifications as a topic map vocabulary (Garshol, 2004).

Dagobert Soergel argued similarly:

> The relationships between document components in a document model, the tags in a document template or a metadata schema, the table structure in a relational database (or the object structures in an object-oriented database), and the relationships between concepts can all be traced back to (or defined in terms of) an entity-relationship model […] Such a model is an ontology, so all structures in a digital library can (and should) be conceived as subsets of an overarching ontology (Soergel, 2009, p. 38).

It therefore seems that all kinds of "indexing languages", "knowledge organizing systems" and classification systems may thus be considered kinds of ontologies. It is a question of whether other kinds of KOS are needed for KO theory or whether other kinds of KOS should just be considered special cases of ontologies with more limited ranges of semantic relations.

However: *whatever KOS we are talking about we face the same kind of fundamental problems: How do we define classes, determine what should be assigned to a given class and how do we determine the relations between classes?*

2. CHALLENGES TO BIBLIOGRAPHICAL CLASSIFICATION IN THEORY AS WELL AS IN PRACTICE

At the practical level, we are facing the challenge of libraries almost entirely ceasing producing/using their own classification systems as well as classifying their books themselves (this is, for example the case in the two largest Danish libraries). The decision to do this may be made based on the following considerations:

1. Many libraries now rely mainly on the *Dewey Decimal Classification (*DDC) made by the Library of Congress (LC) and disseminated in the MARC records rather than making their own classification of each document;

2. Many library directors expect that, in the future, large scanning projects (such as that which is being conducted by Google) may enable full text searches to be carried out of all available content. For this reason, they may consider it a waste of resources to classify or index books;

3. Many libraries now also rely on user tagging and may perhaps expect that this will somehow act as a substitute for professional indexing and classification;

4. Users mostly find the books they need using tools other than the library online public access catalog (OPAC)

At the practical level we see a strong tendency towards centralization. Many rely on *Library of Congress Classifications*.

Given this tendency to rely of Library of Congress, there is a need for examining the quality of LC classification. There is also a need to focus more on bibliographical documentation databases such as MEDLINE.

At the theoretical level we are facing other challenges. We now have Google, for example, which students use far more than they use library catalogs in order to find what they need (cf., de Rosa et al., 2005, 2006; Pors, 2005).

This triggers the question: can information retrieval (IR) theoretically be carried out perfectly well without any kind of "classification" (here understood as metadata assigned by information professionals)? Information scientist Karen Sparck Jones (2005) argued that techniques such as "relevance feedback" would remove the need for classification as it is commonly understood. Previously information scientist Gerard Salton also argued:

> Acting as if we were stuck in the nineteenth century with controlled vocabularies, thesaurus control, and all the attendant miseries, will surely not contribute to a proper understanding and appreciation of the modern information science field (Salton, 1996, p. 333).

I do not doubt that automatic classification is possible and desirable - for many purposes, mind you. What I do claim is that any collection of documents or information can be classified in many ways and that each way may be less fruitful for a given purpose.

Both automatic and manual indexing systems are often supposed to be neutral and the best one for any purpose! It is this positivist epistemology that I believe we need to challenge. We need to prove that Google, for example, is not always good enough.

3. THE FUNDAMENTAL ISSUE: HOW DO WE DISTINGUISH GOOD FROM BAD CLASSIFICATIONS?

What criteria should be used to make classification decisions such as "A belongs to class X"?

We are not used to considering such problems in Library and Information Science and Knowledge Organization.

Often we consider our classifications to be based on "established knowledge," and perhaps we think that it is not our business to examine how this knowledge has been established (or whether it is controversial). My claim is that we cannot avoid considering this fundamental issue.

What are the bases for classifications?

- Some classifications are based on logic (e.g., that even numbers are numbers). The philosophical school of "conceptual analysis" is an attempt to generalize the use of a priory analysis for classification.
- Some classifications are based on empirical studies. A drug is classified as, e.g. tranquilizer, based on medical experiments.
- Some classifications are based on human conventions (e.g. the borders of a country, who is a royal person).
- Some classifications are based on heritage (e.g., who belongs to a certain family)
- Some classifications are based on purpose (e.g. tools for cooking).
- Some classifications are based on a mixture of criteria (e.g., combined logical, empirical and pragmatic criteria)

Given different classifications of a set of elements: How do we determine which classification is best? To evaluate a classification is to consider the methods by which it has been produced and to evaluate the logic, empirical studies, human conventions, the genealogy (in a wide sense of this word), and the goals the classification is meant to serve. To evaluate classifications is - in other words - to engage in the research which lie behind the classification.

Because consensus is seldom, classification involves the considering of different theories, interests and views. Information specialists involved in classification therefore have to consider the different theories, metatheories and "paradigms" in the domain (see Ørom, 2003 as a good example from arts). In chemistry Helium is classified as a Noble Gas. However, Stowe's physicist periodic system (based on quantum mechanics) classifies Helium with the Alkaline Earth Metals (Channon, 2011). Which classification is best? How should information science behave in relation to lack of scientific consensus?

I do not believe that we can get closer than to say that it is important that information scientists know about different theoretical positions and try to argue for the best solution to a given purpose. My basic point is: *We cannot be properly scholarly about classification by neglecting the research on which they are based.*

4. EVIDENCE BASED PRACTICE AS AN ARGUMENT FOR THE NECESSITY OF CLASSIFICATION

Evidence-based practice (EBP) is an influential interdisciplinary movement that originated in medicine as evidence-based medicine (EBM) about 1992. EBP is of considerable interest to library and information science (LIS) because it focuses on a thorough documentation of the basis for the decision making that is established in research as well as an optimization of every link in documentation and search processes. Explicit norms should be made for investigations that are most relevant, and a hierarchy of the value of different kinds of research methods as evidence should be made. At the top of the hierarchy are randomized controlled clinical trials (RCTs). At the bottom is evidence from expert committee reports, or opinions and/or clinical experience of respected authorities.

Between top and bottom is a range of other research methods. A typical hierarchy looks like this:

> Ia Evidence from a meta-analysis of RCTs [randomized controlled trials]
>
> Ib Evidence from at least one RCT
>
> IIa Evidence from at least one controlled study without randomization
>
> IIb Evidence from at least one other type of quasi-experimental study
>
> III Evidence from non-experimental descriptive studies, such as comparative studies, correlation studies and case-control studies
>
> IV Evidence from expert committee reports, or opinions and/or clinical experience of respected authorities (Jainer; Javed; Simpson, 2005, pp. 395-396).

This means that classification of documents by research methods becomes important from the point of view of EBP and EBM has also influenced the indexing of articles in MEDLINE.

There is a huge difference in using such scientific criteria for classifying documents compared to the tradition of "user studies" or "user's view of relevance" in LIS. In that respect EBP is really important for LIS!

However, even if EBP is accepted, could classification - in the meaning of assigned metadata – not be made automatically? Is classification still necessary?

Of course automatic classification of documents may be made on the basis of criteria from EBP. How well such an automatic classification can perform cannot be said in advance. An important point is, however, that it is not an algorithm that determines the criteria of classification. That can only be done by human beings. These criteria are clearly not the ones used by Google. *EBP is therefore a clear demonstration that human negotiated criteria for classification are needed.*

In my paper on EBP (Hjørland, 2011) I have some reservations about this approach based on the philosophy of science. These reservations do not remove the need for classifying documents according to scientific standards. On the contrary: It makes classification less mechanical and formalist. *This is an indication that automatic classification of research methods might be difficult to obtain at a sufficient level of quality.*

5. CLASSIFICATION AND EPISTEMOLOGY

In my opinion it is a mistake that certain epistemologies believe that classifications can be made in neutral ways. Different epistemologies lead to different research methods and different classifications. Questions about human interests, purposes and values are always involved in constructing and evaluating classifications.

- Empiricism is the tendency to emphasize (pure) observations. It is used, for example, in automatic cluster analysis and other numerical taxonomies.

- Rationalism is the tendency to emphasize (pure) thinking (logical analysis and logical division). It is dominating i the facet-analytic approach to classification.

- Historicism is the tendency to emphasize origins of both what is classified and the criteria for classification.

- Pragmatism is the tendency to emphasize goals, values and consequences of alternative ways of classifying.

My pragmatic view may be formulated this way: Experiencing (the ideal method of empiricism) and thinking (the ideal method of rationalism) are two mutually interdependent processes, which, in an iterative historical development, may lead to stable theories and thus to what we regard as the objective reality. In this process, pragmatic factors are also involved (although seldom recognized). The processes are mutually dependent because our observations are theory-dependent and our thinking is dependent on our concepts and thus has a degree of "empirical control".

As formerly stated this view implies that a given thing does not "belong" to a class (and a document does not "have" a subject): Things may be classified in many ways and documents may be assigned different subjects. Different methods and criteria may be used. *In the end it is the purpose of the classification that determines how things should be classified.*

6. CONCLUSION: WHY CLASSIFICATION IS NECESSARY

Bibliographical classification serves activities such as browsing and information retrieval (IR). Criteria of classification are therefore closely related to theories of what is relevant to find or to retrieve. (We also saw that exemplified with EBP). Criteria of relevance are also connected to views of values, goals, interests and consequences.

Sometimes classifications may be based on consensus, but very often classification is involved in different views and metatheories. In order to help people answer questions, finding literature, using terminology etc. information specialists need to know how information/knowledge is organized in domains, genres, "paradigms" etc. To know about, for example, different "paradigms" in arts is important in order to evaluate library classification systems (cf. Ørom, 2003), but also in order to help users find information.

Paradigms and "metatheories" provide basic criteria for what is considered relevant and therefore also for how things should be classified in order to be retrieved or discovered.

The knowledge needed to classify literature is therefore also needed for other kinds of information services, such as information literacy education. This kind of knowledge is demanding and costly, but I see no alternative if classification is to be regarded a scholarly field.

Information science is a metascience related to science studies and related fields. Our core focus is the organization of knowledge for IR in databases and on the web.

7. REFERENCES

AITCHISON, J. "A classification as a source for thesaurus: The Bibliographic Classification of H. E. Bliss as a source of thesaurus terms and structure". *Journal of Documentation*, 1986, v. 42, n. 3, pp. 160-181.

AITCHISON, J. "Thesauri from BC2: Problems and possibilities revealed in an experimental thesaurus derived from the Bliss Music schedule". *Bliss Classification Bulletin*, 2004, v. 46, pp. 20-26.

BLISS, Henry Evelyn. *A system of bibliographic classification.* New York: H. W. Wilson, 1935.

BROADFIELD, A. *The Philosophy of Classification*. London: Grafton, 1946.

BROUGHTON, V. "A faceted classification as the basis of a faceted terminology: Conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification (2nd Ed.)". *Axiomathes*, 2008, v. 18, n. 2, pp. 193-210.

CHAN, Lois Mai. *Cataloging and classification: an introduction*. 2nd ed. New York: McGraw-Hill, 1994.

CHANNON, Martin. "The Stowe Table as the Definitive Periodic System". *Knowledge Organization*, 2011, v. 38, n. 4, pp. 321-327.

DE ROSA, C., CANTRELL, J., CELLENTANI, D., HAWK, J., JENKINS, L.; WILSON, A. *Perceptions of Libraries and Information Resources. A Report to the OCLC Membership*. Dublin, Ohio USA: OCLC Online Computer Library Center, Inc., 2005. Retrieved 2009-10-10 from: *http://www.oclc.org/reports/pdfs/Percept_all.pdf*

DE ROSA, C.; CANTRELL, J.; HAWK, J.; WILSON, A. *College Students' Perceptions of Libraries and Information Resources. A Report to the OCLC Membership*. A Companion Piece to Perceptions of Libraries and Information Resources. Dublin, Ohio USA: OCLC Online Computer Library Center, Inc., 2006. Available at *http://www.oclc.org/reports/pdfs/studentperceptions.pdf* [accessed October 10, 2009]

GARSHOL, L. M. "Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all". *Journal of Information Science*, 2004, v. 30, n. 4, pp. 378-391.

GOLUB, Koraljka. "Automated Subject Classification of Textual Documents in the Context of Web-Based Hierarchical Browsing". *Knowledge Organization*, 2011, v. 38, n. 3, pp. 230-244.

HJØRLAND, Birger. "Concept theory". *Journal of the American Society for Information Science and Technology*, 2009, v. 60, n. 8, pp. 1519-1536.

HJØRLAND, B. "Evidence based practice: An analysis based on the philosophy of science". *Journal of the American Society for Information Science and Technology*, 2011, v. 62, n. 7, pp. 1301-1310.

HODGE, G. *Systems of knowledge organization for digital libraries. Beyond traditional authority files.* Washington, DC.: Digital Library Federation, Council on Library and Information Resources, 2000. Available at *http://www.clir.org/pubs/reports/pub91/contents.html* [accessed May 16, 2011]

JAINER, A. K.; JAVED, M. A.; SIMPSON, I. "Evidence-based practice and its relevance to psychiatry". *Pakistan Journal of Medical Sciences*, 2005, v. 21, n. 4, pp. 395-398. Retrieved December 17, 2010 from *http://pjms.com.pk/issues/octdec05/pdf/evidence_based_practice.pdf*

LANCASTER, F. W. *Indexing and abstracting in theory and practice*. London: Library Association, 2003.

ØROM, A. "Knowledge organization in the domain of Art Studies – History, transition and conceptual changes". *Knowledge Organization*, 2003, v. 30, n. 3/4, pp. 128-143.

PORS, Niels Ole. *Studerende, Google og biblioteker: En undersøgelse af 1694 studerendes brug af biblioteker og informationsressourcer*. Biblioteksstyrelsen og Danmarks Biblioteksskole, København, 2005. Electronic publication, available at http://www.statensnet.dk/pligtarkiv/fremvis.pl?vaerkid=45550&reprid=1&iarkiv=1 [accessed May 16, 2011]

RIESTHUIS, G. J. A.; BLIEDUNG, St. "Thesaurification of the UDC". *Tools for knowledge organization and the human interface.* Index Verlag, Frankfurt, 1991, v. 2, pp. 109-117.

SALTON, G. "Letter to the editor. A new horizon for information science". *Journal of the American Journal for Information Science*, 1996, v. 47, n. 4, p. 333.

SOERGEl, D. "Digital Libraries and Knowledge Organization". In: Kruk, S-R.; McDaneil, B. (Eds). *Semantic digital libraries*. Berlin: Springer, 2009, pp. 9-39. Available at *http://www.dsoergel.com/NewPublications/SoergelDigitalLibrariesandKnowledgeOrganization.pdf* [accessed May 16, 2011]

SPARCK JONES, K. "Revisiting classification for retrieval". *Journal of Documentation*, 2005, v. 61, n. 5, pp. 598-601.