Theories of knowledge organization — theories of knowledge

Birger Hjørland

Keynote March 19, 2013

13th Meeting of the German ISKO (International Society for Knowledge Organization),

Potsdam, 19th to 20th March 2013

Abstract

Any ontological theory commits us to accept and classify a number of phenomena in a more or less specific way – and vice versa: a classification tends to reveal the theoretical outlook of its creator. Objects and their descriptions and relations are not just "given" but determined by theories. Knowledge is fallible and consensus is rare. By implication, knowledge organization has to consider different theories/views and their foundations. Bibliographical classifications depend on subject knowledge and on the *same* theories as corresponding scientific and scholarly classifications. Some classifications are based on logical distinctions, others on empirical examinations, and some on mappings of common ancestors or on establishing functional criteria. To evaluate a classification is to involve oneself in the research which has produced the given classification. Because research is always based more or less on specific epistemological ideals (e.g. empiricism, rationalism, historicism or pragmatism), the evaluation of classification includes the evaluation of the epistemological foundations of the research on which given classifications have been based. The field of knowledge organization itself is based on different approaches and traditions such as user-based and cognitive views, facet-analytical views, numeric taxonomic approaches, bibliometrics and domain-analytic approaches. These approaches and traditions are again connected to epistemological views, which have to be considered. Only the domain-analytic view is fully committed to exploring knowledge organization in the light of subject knowledge and substantial scholarly theories.

1. Ontological commitment

Knowledge organization (KO) is about classifying knowledge, for example, to define concepts and determine their semantic relations, i.e. to define "cat" (*Felis catus*) and its relation to other concepts such as "mammal" (*Mammalia*) (in this case the semantic relation is termed an "is-a" relation, a "generic" relation, a "genus-species" relation among others). In other words: KO is about concepts and their semantic relations (and at the same time about the real world, here: animals).

How do we know what a cat is (i.e. what the concept "cat" means)? How do we know the relation between "cat" and other species (such as "dog")? How do we know what "a species" means? And how do we know the relations between a given species and genera, families, kingdoms etc.? These are far less trivial questions than most people believe them to be: in mainstream biological systematics major groups of animals (such as fishes and reptiles) are no longer regarded as valid taxa (i.e. groups of organisms recognized as formal units, although they continue to be studied and written about), cf. Blake (2011, 467). This example also shows that terms and classifications (such as "fishes" and "reptiles") are inconsistently used even within one domain (biology): the new taxonomic victory has been incomplete.

Normally non-experts would just say that **we know** what a cat is and that we know that it is a mammal. If challenged we might look it up in an authoritative source, either a general encyclopaedia like *Encyclopaedia Britannica* or an authoritative biological handbook (such as Wilson and Reeder 2005), or ask some experts. But of course, different sources may disagree and in the end we have to argue why the chosen source is authoritative. If we take the question to the extreme we have to leave second-hand knowledge (Wilson 1983) and involve ourselves in research in biological taxonomy and the philosophy of classification.

Many influential philosophers subscribe to the principle of fallibilism, which is a philosophical doctrine, most closely associated with Charles Sanders Peirce, which maintains that our scientific knowledge claims are invariably vulnerable and may turn out to be false. Fallibilism does not insist on the falsity of our scientific claims but rather on their tentativeness as inevitable estimates; it does not hold that knowledge is unavailable, but rather that it should always be considered provisional (Rescher 1998). We have "known" for a long time that the planets of our sun are: Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune and Pluto. In August 2006, however, the International Astronomical Union redefined the term "planet", and classified Pluto along with some asteroids as a dwarf planet. This example thus confirms the principle of fallibilism (and by implication all knowledge organization systems (KOSs) had to be updated). This is also the case with the classification of animals:

> Scientists aim to describe a single "tree of life" that reflects the evolutionary relationships of living things. However, evolutionary relationships are a matter of ongoing discovery, and there are different opinions about how living things should be grouped and named. EOL [Encyclopedia of Life] reflects these differences by supporting several different scientific "classifications". Some species have been named more than once. Such duplicates are listed under synonyms. EOL also provides support for common names which may vary across regions as well as languages http://eol.org/pages/2850509/names.

By implication it is not wise to claim that "we know X to be a kind of Y" or that "we know that concept X is semantically related to concept Y by a certain relation such as a genus-species relation". It is wiser to say "based on current theory X is considered a kind of Y". We then have to examine whether or not there is scientific consensus. Non-specialists tend to overestimate the degree of consensus in science, as pointed out by Broadfield (1946, 69-70): "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."

In cases where there is no consensus the classifier has to make a decision based on an evaluation and negotiation of the different positions. Such a classification cannot be neutral, but will favour some views at the expense of others. This has been clear for a long time and also expressed in my former publications. Feinberg wrote, however:

> While Hjorland (1998[b]) then asserts that classification is not neutral and is theory-laden, this seems to be based more on the idea that the material to be classified is theory-laden, than that [a] classificationist is actively designing a certain view in the classification. A domain, for

> example of psychology, exists; it seems to be the classificationist's job to find and describe it, not to define or build it (Feinberg 2008, 19-20).

This quote does *not* reflect my opinion as stated in my former writings. Hjørland (1992, 189) concluded: "Thus an analysis of a subject is itself, at its most profound, a part of the scientific process of knowledge gathering" (implying that the classificationist's job is not neutral). This was correctly understood and referred to by Melodie Fox:

> Hjørland (2008[b], 335), on the other hand [contrary to Rick Szostak], believes that "'neutrality' and 'objectivity' are not attainable" and that "Any given classification will always be a reflection of a certain view or approach to the objects being classified" whether it is easily detectable or not (Fox 2012, 302).

Feinberg also seems to recognize this in the following quote: "It seems to me, though, that Hjorland's case study of subject analysis, in which he determines the subject of a psychology book, depends on a quite particular viewpoint or theory of psychology" (Feinberg 2008, 73). Yes indeed: classifications are theory-dependent and thus not neutral. I thought we agreed on this? Why then this objection? The main difference between my view and Feinberg's is probably that I recognize that the criteria that are relevant for the classificationist are not just his or her private criteria, but usually are related to or derived from theories which tend to be publicly shared as "paradigms". Therefore classification supposes subject knowledge (the ability to critique different subject theories and their ideological impact on classifications).

We cannot – as classification theorists – say which view should be preferred in matters of scholarly controversy (although we may have our private assumptions or preferences). This condition may be the reason for Feinberg's (2008, 277) complaint about Hjørland's domain-analytic view that "[t]he basic construct of domain is not concretely defined, for example, which makes it difficult to determine how to set boundaries for analysis". My answer is that such boundaries cannot be set up a priori, and that they are always provisional; all we can say is that the best qualified decision is one based on the best understanding of the scholarly evidence as well as insight into the implications of the alternatives, and into pragmatic and ethical issues (Blake 2011, 469; Mai 2012). In other words: the classifier must be qualified to discuss the different views, he or she must be meta-theoretically well informed. Feinberg here seems to demand a theory-independent classification, which is in contrast to my (and to her own) claimed position.

The relation between theories and classifications leads to the notion of ontological commitment:

> The notion of ontological commitment has come to prominence in the second half of the twentieth century, mainly through the work of [Willard Van Orman] Quine [1908-2000] […]. On Quine's view the right guide to what exists is science, so that our best guide to what exists is our best current scientific theory: what exists is what acceptance of that theory commits us to (Craig, 1998).

Of course, classifications are not always scientific (or scholarly). We also have everyday classifications of, for example, pets and aquarium fish, kinds of clothes, administrative rules and much else. Anybody is

allowed to classify animals by their colours, "sweetness", size or any other criteria relevant for a particular situation. However, if our KOSs should support persons to have what we (following Wilson 1968, p. 21) may call the best textual means to their ends, then KOSs have to be based on some functional criteria. Often the general language contains functional criteria different from scientific language. Such differences are explored in – among other fields – sociolinguistics, where the functions of different concepts and distinctions for different groups of people have been explained functionally (Ammon 1977). Science and scholarship should be considered one among other kinds of discourse communities developing their own pragmatic conceptual structures. And of course, new kinds of classifications are being developed all the time (e.g. in books about animals for children, in creative museums etc.) The point is, however, that whatever domain is in need of *professional* information services and therefore knowledge organization systems developed within our field should be explored from its ability to serve its target group or its ideal purpose. Epistemological analysis is part of domain analysis and is not just about science, but also about everyday knowledge. Mainstream scientific psychology may, for example, be criticized for downgrading personal experience and the kind of knowledge achieved through the arts. But to make that argument and to design a classification system accordingly requires scholarly arguments. The point is also that KO as a field cannot serve classifications where there are no criteria to decide whether one system is better than another, and no goal at all to fulfill (as Feinberg 2008, 6, seems to believe).

In conclusion: Any ontological theory commits us to identifying and classifying a number of phenomena in a specific way – and vice versa: a listing and classification of a number of phenomena may reveal the theoretical outlook of its creator ("show me your classification and I'll tell what theory you subscribe to"). Not every scientific theory may imply different ontologies, however. The competing theories that global warming is caused by human activities versus by activities on the sun may both share the overall understanding of what phenomena exist and their relations. Ontological theories are theories that imply claims of the things that exist in a domain (such as cats, fish and planets, atoms, antimatter, information or information needs) – and such theories are mostly considered *fundamental* scholarly theories or "paradigms".

2. Scientific versus bibliographic classifications

Mai (2004, 41) argued that "[s]cientific classification and logical division has worked fairly well in the classification of natural kinds, such as Linnaeus' classification of living things" (a challenge of the view that logical division works well in the classification of living things is given in Hjørland 2013a). Mai continues (p. 42): "It is my contention that scientific classification of natural objects, and the bibliographic classification of the content of a document, are distinct for two main reasons. The first has to do with when and how the items are classified, and the second has to do with the nature of the classified items." I disagree with this statement (as discussed in Hjørland 2008a). I find Mai's understanding harmful because it undermines the important relation between subject knowledge and bibliographical classification (e.g. between knowledge about zoological taxonomy and the design of classification systems on animals for bibliographic databases). For a qualified and relevant description of the relation between biological taxonomy and bibliographical classification see Blake (2011).

Blake (2011) writes that cladistics is a novel classificatory method and philosophy adopted by zoologists in the last few decades, which has provided a rather turbulent state of zoological classification. He writes:

> [Z]oologists see biological classification as both an expression of theories about the relationships between taxa and as an information storage and retrieval system. Mayr (1982, 240-1) argues that the second of these functions imposes limits on both the number of taxa a higher taxon can sensibly contain and on the number of levels appropriate in a hierarchy. Thus cladistics, with its deep hierarchies, can be seen as a move towards greater scientific accuracy at the expense of efficient information retrieval. This inefficiency with regard to information retrieval helps explain why many monographs and other publications continue to organise their material using Linnaean ranks rather than hierarchies of clades (Blake 2011, 466).

> At present, many, perhaps most, current bibliographic classifications for mammals reflect quite outdated science. The latest edition of DDC, for example, arranges mammals in essentially the same way as the second edition of 1885. Revisions since DDC2 have mainly focused on adding detail and giving more guidance to users about where to place certain taxa. New (1996) and New and Trotter (1996), in their accounts of the changes introduced to the zoology schedule in DDC21, emphasise pragmatic concerns such as avoiding the re-use of numbers, rather than keeping up with developments in zoology. Indeed, some of the changes made in DDC21, such as moving the monotremes to a position between the marsupials and placentals (Mitchell 1996, 1181), represent a move away from scientific accuracy in the interests of practical concerns such as the efficient use of notational space. Such "outdated" classifications may still do their job well. The library of the Zoological Society of London uses its own scheme, devised in the 1960s and largely based on the Bliss Bibliographic Classification, to classify the monographs it holds. The librarian reports that, in most cases, her patrons are able to retrieve items and browse the collection effectively (Sylph 2009) (Blake 2011, 469-470).

Blake also refers to a text about forthcoming revision of the UDC:

> UDC schedules have used the Linnaean system from its first editions, and through this revision, this classification structure will be preserved. But, since the growing presence of Cladistics in academic sources cannot be ignored, some of its less controversial elements will be incorporated. By doing this, UDC systematics sections will benefit from the best of both classification currents, carefully avoiding the existing problems and conflicts (Civallero 2011, 10).

Blake and Civallero thus express the view that classification of natural objects is also subject to the same kinds of theory dependence, interpretation and difficulties as documents are.

Blake also claims that the aim of biological theories and the aim of classification for information retrieval may be in conflict. He even claims that "'outdated' classifications may still do their job well". Can that really be true? If it is true, might the reason be that library classifications do not serve *advanced* retrieval purposes (within front-end research or that libraries and databases do not support the dissemination of

new knowledge to the general public)? If we have such a low level of ambition concerning classification systems is there then a need for KO as a scholarly research discipline? We are here dealing with three levels: front-end biological research using new classifications, mainstream biology being in a process of catching up and still *also* using some obsolete classifications, and information science standing in a conflict between advanced theory and literary warrant (because much of the literature to be classified is written from obsolete positions).

Another indication of the coherence between the classification of objects and documents is Anders Ørom's description of how different "paradigms" in art studies influence how literary works are organized, how art exhibitions are organized and how library classification systems are organized.

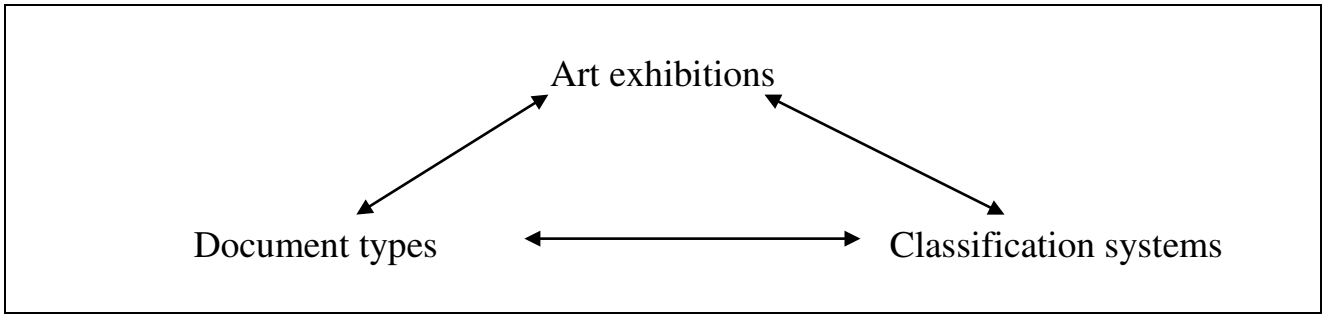

Figure 1
Social values – world views – scholarly paradigms
(After Ørom 2003, 132)

3. The epistemological basis of classifications

Classifications have different bases:

- 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 priori analysis for classification (Hanna 1998).
- Some classifications are based on empirical studies. A drug, e.g. a tranquilizer, is classified as 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). The so-called cladistics school in biological systematics which today is the dominant school is based on this principle (this is also based on empirical research, but not on the doctrine of empiricism).
- 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, historicist and pragmatic criteria).

Logical, empirical, historicist and pragmatic methods may each have applications for which they are especially relevant but each may also be generalized and used more widely because of traditions and ideologies.

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, knowledge of 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 lies behind the classification in order to check its validity.*

All research is influenced by epistemological norms or commitments. There is no simply "correct way of doing research" or one correct and all-encompassing "scientific method", and also in the theory of knowledge consensus is seldom. In my view, versions of pragmatism/activity theory are the best candidates for fruitful philosophy of enquiry, but this issue is still open and is today in a somewhat confused condition. The classical theories of empiricism and rationalism are still very much alive and influential in contemporary research (although mostly unrecognized). These theories have been characterized as a trap (Mammen 2008, 25; Toulmin 1999), and the point here is that if we understand their shortcomings, we may avoid the trap and do better research leading to better classifications. Empiricism and rationalism used to be considered the fundamental epistemological positions (and their combination was tried by the logical positivists at the beginning of the 20th century without success). Because of their shortcomings, we need to include some alternatives. I consider four theories as the basic epistemological theories: empiricism, rationalism, historicism and pragmatism:

Simplified relevance criteria in four epistemological schools			
Empiricism	*Rationalism*	*Historicism*	*Pragmatism/activity theory*
Relevant: Observations, sense data. Induction from collections of observational data. Intersubjectively controlled data. Non-relevant: Speculations, knowledge transmitted from authorities. "Book knowledge" ("reading nature, not books"). Data about the observers' assumptions and pre-understanding.	Relevant: Pure thinking, logic, mathematical models, computer modelling, systems of axioms, definitions and theorems. Low priority is given to empirical data because such data must be organized in accordance with principles which cannot come from experience.	Relevant: Background knowledge about pre-understanding, theories, conceptions, contexts, historical developments and evolutionary perspectives. Low priority is given to decontextualized data of which the meanings cannot be interpreted. Intersubjectively controlled data are often seen as trivia.	Relevant: Information about goals and values and consequences both involving the researcher and the object of research (subject and object). Low priority (or outright suspicion) is given to claimed value-free or neutral information. For example, feminist epistemology is suspicious about the neutrality of information produced in a

			male-dominated society.

This classification of epistemological theories is, of course, an abstraction and simplification of the wide range of concepts and viewpoints used today, e.g.:

- actor-network theory
- cognitivism
- critical rationalism
- critical realism
- critical theory
- dialectical materialism/Marxism
- feminist epistemology
- hermeneutics
- paradigm theory
- phenomenology
- postmodernism (late modernism)
- semiotics
- social epistemology
- social constructivism

All of these and more are theoretical positions which may be encountered in information science and KO. It is not, however, always clear in what way each position makes a difference for, say, classifying animals (or anything else). If somebody works from a specific epistemological position, that person should be able to say in what manner this position makes a difference for the specific work. If this cannot be done, the position is of no consequence (and trivial). From my own work I have found that the above-mentioned four positions are able to catch most of the important implications in the positions listed. There is, however, a big need for some consensus in KO about epistemologies and their implications.

The most important issues in the criticism of empiricism, rationalism and "positivism" probably are the neglecting of knowledge as a social and historical product made to serve certain purposes and interests and by implication the need reconsider knowledge claims in the light of new purposes, conditions and interests.

> Textbooks like Harter (1986), Lancaster (2003), Large, Tedd & Hartley (2001), and Svenonius (2000) can be characterized as texts that solidify the use of technical and managerial language in LIS in the sense that they are basically how-to books, constantly referring to techniques, standards, principles, methods and rules. If one's professional knowledge base has such texts at its foundation, no critical attitude is developed nor demanded because these textbooks do not question at all the role of information seeking or of knowledge organization systems in culture and society. They do not provide students with a language, an understanding, a knowledge that make them capable of participating in public discourse debating the functionality and legitimacy of these systems" (Andersen, 2005).

I believe Jack Andersen's quote can be interpreted as a critical epistemological view of KO. (And, by the way, he is inspired by activity theory).

A critical view cannot, however, be separated from knowledge about technical aspects of retrieval systems. There is a need to revise theories of KO, not just to replace them with critical attitudes. In the rest of this proposals for such revisions are put forward.

4. Approaches to knowledge organization

KO is a field about classifying and indexing documents, for example biological documents. As such, it needs to consider the fundamental theories and paradigms in the domains which it organizes (as we saw above: it needs to consider, for example, the new cladistics paradigm in biological taxonomy). But it is also itself a field influenced by different paradigms related to theories of knowledge. In this paper a short outline is given with references to more detailed treatments in other papers, published or in the pipeline.

4.1 Automatic versus human classification

In overviews of KO a fundamental difference between computer-based versus human-based classification and indexing is often made (e.g. Anderson and Pérez-Carballo 2001a+b). In Hjørland (2011) I argue, however, that this distinction theoretically is unfruitful. One argument is that humans may use primitive rules and in reality function as a computer when indexing documents. Human beings classify according to what they have learned or been instructed to do, or how they believe they should do the indexing; computers likewise classify according to the techniques and views which were available to their programmers at the time the programming was done. Both humans and computers thus index in very different ways based on different views, which at the deepest level are related to ontological and epistemological views. Alternatively I therefore suggest that both human and computer indexing may index in accordance with one or other of the basic theories of knowledge (empiricism, rationalism, historicism and pragmatism); these epistemologies are fundamental theories of KO (see further below; see also Hjørland 2011).

4.2 User-based and cognitive classifications

User-based and cognitive views have been influential since the 1970s. Hjørland (2013b) is a critical analysis of this approach. With Hansson (2006, 33) I find that "In knowledge organization theory, cognitive perspectives have not been as dominant as in information behavior research. The reason for this is it is practically impossible, at least in the long run, to avoid connecting knowledge organization and classification research to the actual content of the documents and document collections in relation to the classification and indexing performed. This can seem trivial, but it is actually not". The basic issue in KO is about questions such as: Should document A be classified in class X? Is term A synonymous with term B? User-based and cognitive approaches are not appropriate ways to answer such core issues in KO. The tendency to ask users is seen as a kind of positivism in which the empirical studies of users are considered better research than the scholarly studies of knowledge domains. The belief that cumulation of empirical data about users may in itself turn out to be useful for classification is seen as a problematic assumption related to empiricism. The user-based tradition thus represents one among other examples of how empiricism as a theory of knowledge has influenced KO.

4.3 Facet classifications

Hjørland (2013a) found that the facet-analytic approach is based on the epistemology of rationalism. The strength of this approach is its logical principles and the way it provides structures in knowledge organization systems (KOSs). The main weaknesses are 1) its lack of empirical basis and 2) its speculative ordering of knowledge without basis in the development or influence of theories and socio-historical studies. It seems to be based on the problematic assumption that relations between concepts are a priori and not established by the development of models, theories and laws. This tradition thus demonstrates how rationalism as a theory of knowledge has influenced KO.

4.4 Numeric taxonomic approaches

Statistical methods such as cluster analysis, factor analysis etc. are used in many different sciences and on many different kinds of data (e.g. for classification of diseases or biological organisms). They are also used for classifying documents (vector space models, latent semantic indexing etc.) and may therefore be considered an approach to KO. This is an extremely wide and complex field and it may seem hasty and problematic to go into this field in such an overall way that is here attempted. However, these techniques are competing with other approaches to KO (and seemingly have much more success and authority in academia today). I therefore feel that we in KO have to take numeric taxonomic/IR approaches very seriously, and if we want to make room for other approaches, we have to provide convincing argumentation about the limits of the approaches that are competing with the ones we want to defend.

Ellis et al. (1993) provide an overview and a discussion of a broad variety of similarity coefficients in the use of the degree of similarities between objects that contain textual information such as documents, paragraphs, index terms or queries. Their Table 2 lists 27 such measures (classified as distance coefficients, association coefficients and correlation coefficients). However, often coefficients are equivalent or monotonic with each other, which means that it can be shown that the ranking of all measurements between pairs of objects in a specific set is the same using one coefficient as it is using the other. In many cases, however, the different coefficients classify in different ways. Which coefficient should be used in order to measure the similarity between two objects (e.g. between a query and a document)?

> Presented in these terms, the history of research into the use of similarity coefficients in text retrieval appears to betray a lack of progress (Ellis et al. 1993, 141).

The authors refer to critical voices:

> Even in the field of numerical taxonomy, where the use of similarity coefficients has been even more widespread than in information retrieval, Jackson, Somers and Harvey (1989) were moved to conclude that "the choice of a similarity coefficient is largely subjective and often based on tradition or on *a posteriori* criteria such as the 'interpretability' of the results", and went on to quote Gordon (1987): 'Human ingenuity is quite capable of providing a *post hoc* justification of dubious classifications'" (Ellis et al. 1993, 144).

How can progress be made? What epistemological issues are involved? If we consider animals as objects, a property of an animal may be that it has a beak and more specifically that it has a rounded beak. Some schools of biological taxonomy classify animals on the basis of such characteristics (while cladistics classifies solely on the basis of a common ancestor). The school of numeric taxonomy would classify animals on the basis of as many properties of this kind as possible and then use some kind of similarity coefficient to classify similar animals. The empiricist philosophy is committed to the selection of such properties on a basis which is not "biased" by the researchers' selections or theories. However, what the numeric taxonomist has to work with are *the descriptions* of the objects made by themselves or by other (former) researchers. If we assume that no description of an animal can ever be complete, and if we assume that the way researchers describe the properties of animals is informed by their assumptions of what is relevant to describe, then we have in principle just a set of biased descriptions which can be used by similarity coefficients (they may be biased, for example, by giving priority to structural properties rather than ecological properties, to macro properties rather than micro properties etc.). *Taxonomists do not have direct access to the animals themselves, only to sets of descriptions that are in principle always biased.* Such a biased set of descriptions can be more or less homogenous or represent a merging of different priorities of description (Hjørland 1998a, 2008b). The point here is that in order to apply or interpret the results of similarity coefficients we have to give up the empiricist doctrine of "non-biased" descriptions (and collections of such). If we assume that cladistics taxonomy is the best scientific evidence about the classification of animals, then the descriptions and properties from cladistics research should be considered the best (and a kind of reference or standard). And the similarity coefficient that best reflects the cladistics order should be preferred (and considered the norm). We are doing exactly the opposite of the empiricist commitment: we take the theories, not the observations, as our point of departure (but of course, observations form important parts of our theories).

On the next level, we are not dealing with animals, but with documents about animals and animal properties. *Here we have exactly the same epistemological problems.* The assumption that two documents are related with regard to subject if they share the same statistical distribution of words is often held in the tradition of IR. That this assumption may be problematic is easy to demonstrate because two documents in different languages (English and Danish) may be about the same subject matter in spite of their difference in words. Also, from an epistemological point of view, two objects are not just more or less similar, but they are always similar in some respects but dissimilar in other respects. In order to identify documents by algorithmic means, we need a set of criteria for how relevant documents can be distinguished from non-documents. The mainstream tendency has been either to apply "largely subjective [criteria] and often based on tradition or on a posteriori criteria such as the 'interpretability' of the results" or to seek such criteria in the mind of the users. Alternatively I have suggested that scientific, scholarly and epistemological criteria are what should be preferred (Hjørland 2010, 2013b). For example, two documents may be considered related if they are about the same organism or taxon as described by *current* biological theory. From another perspective the same documents may be relatively unrelated.

The empiricist doctrine of non-biased descriptions of documents is non-tenable (this goes for the use of descriptors, titles, text or bibliographical references and any other element or combination thereof). Any choice will make a difference with regard to the classification of documents, and how can we decide which choice is best? Well, if we assume that cladistics taxonomy is the best scientific evidence about the

classification of animals, then this theoretical view should also inform our evaluations of document descriptions and similarity measures.

It may be common knowledge that numerical taxonomy approaches require substantial theoretical knowledge. This is strongly emphasized by Hetherington (2000, 40ff.). He refers to Kaplan's (1964) "law of the instrument" as the problematic tendency to use techniques, not theory, to direct scientific practice. Theory should be used in the research process to establish guidelines for data analysis.

> Although multivariate statistics can generate an impressive array of information, they may nevertheless produce nothing more than "well-dressed" GIGO [Garbage In-Garbage Out] without the guidance of substantive theory (Hetherington 2000, 40).

In the field of classification of mental diseases, Cooper (2005) concluded that one cannot select empirical variables for numerical techniques for classification without a basis in domain-specific theory.

This has also been emphasized in bibliometrics:

> The quality of a SOM map [self-organizing map] or an MDS [multidimensional scaling] map should be evaluated by experts in the area studied, as no objective means exist for assessing unknown domains. This opinion is shared by Tijssen [1993], […] he [Tijssen, 1993] offers empirical data to show that the cognitive perception of a group of experts in one subject area with respect to the same map can be very diverse (Moya-Anegón et al. 2006, 72).

In spite of these many expressions about the necessity of substantial theory, such theory seems to be missing in the literature on information retrieval. The overall tendency in IR research and numeric taxonomy has been committed to the empiricist ideal. Mainstream IR therefore – as user studies – represents an example of the influence of empiricism in KO.

4.5 Bibliometric classifications

In Hjørland (submitted) I make the distinction between KOSs reflecting intellectual KO and KOSs reflecting social KO:

- The intellectual aspect of KO is knowledge organized in concepts, propositions, models, theories and laws. Such intellectual organizations are primarily structured via relations of explanatory coherence (Thagard 1992, 9), which are again primarily related to questions concerning truth.
- The social aspect of KO is knowledge organized into academic departments, disciplines, cooperative networks, administrative bodies etc. Such social organizations are primarily structured by the social division of labour in societies, which is again primarily related to questions concerning social relevance, authority and power.

We thus have two kinds of KO driven by criteria which may support or oppose each other in complex mutual interactions. Sometimes there are agreements between intellectual and social organizations. In biology, for example, "mammals" is a theoretical concept in taxonomy and the American Society of Mammalogists is a social organization. Often, however, there are disagreements. Paleontology, for

example, is a discipline studying prehistoric life. “Prehistoric species” is not, however, a concept in biological taxonomy in the same way that “mammals” is:

> The division of the Tree of Life into an “extinct” and a “living” section is an artificial approach based on a disciplinary point of view which does not work well for systematics. Based on this approach, for instance, a group [such] as Mammals is divided into organisms studied by Palaeontology (species known through their fossil remains) and those studied by Zoology (living and recently extinct species). The division would not cause any issues if it would be a simple placing of strictly extinct classes into Palaeontology schedules and strictly living classes in Zoology. But many animal and vegetal groups have both extinct and living species, and therefore, they should be present in both schedules. Up until now, the practical solution provided by UDC was the use of parallel divisions: taking Zoology tables as the reference model, Palaeontology can be subdivided in parallel (Civallero 2011, 14).

In general, it cannot be expected that methods based on citation analysis are able to produce intellectual maps such as geographical maps, biological taxonomies or periodical systems.

> A geographical structure, for example, places different regions in a structure that is autonomous in relation to the documents that are written about those regions. You cannot produce a geographical map of Spain by making, for example, bibliometric maps of the literature about Spain [Yet such autonomous structures as maps of Spain are often very useful for information retrieval about Spain] (Hjørland 2002b, 452).

Methods based on citation analysis (e.g. co-citation analysis and bibliometric coupling) represent social organizations and cannot as such be expected to correspond fully to theory-based KOS. They are valuable but cannot substitute domain-analytic studies concerned with substantial theory. Many bibliometric studies are close to mainstream IR and are committed to the empiricist epistemology. However, the understanding in the following quote may provide an alternative:

> [c]o-citation patterns change as the interests and intellectual patterns of the field change (Small 1973, 265).

This understanding opens the doors to a historicist and social epistemology which considers the relation between papers and concepts in the light of research traditions and paradigms. In this way bibliometrics may provide KO with a new and valuable epistemological perspective.

Bibliographic methods cannot render subject knowledge superfluous (but is itself – like numeric taxonomy – dependent on subject knowledge). Although bibliometrics is often associated with domain-analysis, I here argue for considering these approaches separate.

4.6 Domain-analytic classification

The domain-analytic view first of all recognizes the need for subject knowledge in classification and indexing. A fine domain-analytic study is Blake (2011) who demonstrates solid knowledge about zoological

taxonomy and the competing approaches in the field (cladistics, evolutionary taxonomy and the Linnaean system). He also carefully discusses the relations between scientific theory, quasi-taxonomic groupings, and the specific demands that information retrieval puts on classifications (including the principles of literary warrant). Finally, the paper describes the classifications used by biologists in their writings (monographs) and reveals the tendency to use conflicting or inconsistent classifications (corresponding to Ørom's (2003) concept "bricolage").

Another example is Ørom (2003) who outlines different paradigms in the field of art studies and demonstrates the relation between library classifications and art paradigms. A given paradigm reveals itself in the way books of art history are organized, in library classification systems and in the way art exhibitions are organized.

The last example here will be Hjørland (1998b, 2002a), which discussed problems in the classification of psychology. It is demonstrated that there is no consensus about the basic concepts of psychology, but a number of competing schools or "paradigms" each implying its own classification of the field. These schools can – in a philosophical analysis – be related to different basic epistemologies (empiricism, rationalism, historicism and pragmatism). Classifying psychology is not (as claimed by Feinberg 2011, 19) "the union of approaches used to study it". On the contrary, a classification is a subjective choice or negotiation between different views. The difference between a good and a bad classification is that the good classification reveals deep insight concerning the possible choices and dilemmas and is well argued (and has considered counterarguments, including potential counterarguments).

Tennis (2003) asked: "What is a domain?" The answer is that for any specific domain (say, information science) there are conflicting views of how to delimit the field (to say, for example, which journals belong to the field). In a way, information science is something existing to be described. But in other ways it is something that we are in the process of constructing – from our different perspectives and interests. The way we classify a domain is not "objective" but is inevitably "biased" by our interests and perspectives. In my opinion, we cannot and shouldn't (as Tennis demands) make "an operationalized definition, a transferable and standardized definition" of information science: that would be the rationalist approach (practised in facet analysis), which ignores the historical, social and political issues in defining the field. When Carl Linnaeus wrote *Systema Naturae* (1735) botany and zoology were seen as two separate domains. It was the invention of the microscope (and a hundred years of using it) that led to the discovery that all plants and animals consist of cells, which led to the unified domain: biology. How could a classifier define biology without this knowledge? Domains are thus constructed dynamically and cannot therefore a priori be given "an operationalized definition, a transferable and standardized definition". Defining and classifying a domain is therefore best described in terms of the hermeneutic circle. Hjørland and Hartel (2003) described domains as complicated interactions between three facets: (1) ontological theories and concepts about the objects of human activity; (2) epistemological theories and concepts about knowledge and the ways to obtain knowledge, implying methodological principles about the ways objects are investigated; and (3) sociological concepts about the groups of people concerned with the objects.

In the literature of LIS, semantic relations (as displayed in classification systems and thesauri) are sometimes termed a priori relations (Svenonius 2000, 131; Will 2008; ISO 2788 1986, 1; ISO 25964-1). Willpower information, for example, defines:

> Paradigmatic relationship (use for a priori relationship; semantic relationship): Relationship between concepts which is inherent in the concepts themselves. Such relationships are shown in a structured vocabulary, independently of any indexed document (Will, 2009).

This is a problematic terminology: the typical meaning of "a priori" in philosophy as well as in general language is "non-empirical"(Moser 1998), and in most cases it is simply wrong to consider semantic relations as non-empirical. To classify a cat as a mammal is based on the empirical examination of cats based on some criteria (e.g. that mammals are vertebrate animals which feed their young on milk produced by mammary glands. The knowledge that cats feed their young this way is, of course, empirically established). By considering semantic relations in KOSs "a priori" one therefore fails to recognize that the classificationist's job cannot rely just on common sense but has to consider the available evidence. Such relations are often (but not always) determined by scientific research.

The domain-analytic approach to KO is thus the only one which is fully committed to exploring knowledge organization in the light of subject knowledge and substantial scholarly theories. All the other approaches can be understood as attempts to avoid considering the necessary subject knowledge. From the perspective of domain-analysis, such neglect must inevitably lead to a lack of progress.

5. Conclusion

The necessity for subject knowledge in KO (as in the broader field of information science/library and information science) is certainly not a new idea. This kind of knowledge has always been assumed in high-standard libraries and bibliographical databases such as the National Library of Medicine and the MEDLINE database (in parallel with teaching qualifications: the higher the level of teaching, the bigger the demands on subject knowledge).

Also, voices in the research literature of KO have expressed the need for subject knowledge. Richardson and Bliss, for example, considered the implications of the need for subject knowledge for education in librarianship and IS:

> Again from the standpoint of the higher education of librarians, the teaching of systems of classification . . . would be perhaps better conducted by including courses in the systematic encyclopedia and methodology of all the sciences, that is to say, outlines which try to summarize the most recent results in the relation to one another in which they are now studied together (Richardson, quoted in Bliss 1935, 2).

A recent voice is that of Jennifer E. Rowley and John Farrow:

> In order to achieve good consistent indexing, the indexer must have a thorough appreciation of the structure of the subject and the nature of the contribution that the document is making to the advancement of knowledge (Rowley and Farrow 2000, 99).

In spite of such voices, subject knowledge has been and still is extremely neglected in KO. Among the reasons is that information scientists used to be scientists (e.g. chemists) specializing in information science, but it has been difficult for schools of library and information science to attract scientists (and other scholars). Another strong reason may be the feeling that if information science and KO are independent disciplines, they have to have knowledge of their own, not just be based on knowledge from other fields. However, KO is a metascience and is dependent on substantial, domain knowledge. KO and information science in general share with other metadisciplines, such as the philosophy of science, the sociology of science and the history of science, dependence on subject knowledge and at the same time a unique focus.

My claim is that the neglect of the importance of subject knowledge has brought forward a crisis in KO, and that no real progress can be observed in the field. Of course, there is plenty of progress in the development of digital technologies which enable better kinds of knowledge representation and information retrieval. But such progress is brought to us from the outside; it is not something the field of KO has provided. It is important to realize that there is a need to make sure that the KOSs developed or studied within our field are sufficiently based on and related to current scientific theory (that is also the case with approaches based on numeric taxonomic methods). There is no short cut via user studies, common sense or anything else.

Where does this place the theory of knowledge in KO? The first thing to say is that you cannot classify domains on the basis of theories of knowledge (or other metadisciplines, including genre studies, the sociology of knowledge etc.): our studies have to be based on concrete domains. Epistemology is, however, the best *general background* it is possible to teach people within in information science. It is the best general preparation we can provide for people in order to study any domain. The same kinds of philosophical problems seem to show up in all domains, and if the limitations of a certain position have been understood in one domain, it is probable that the same position can also be turned down in another domain. A general lesson from epistemology is that knowledge is created by humans for some specific purposes and serves some interests better than others. Concepts and semantic relations are not a priori or neutral, but should be examined in relation to their implications for the users they are meant to serve.

References

Ammon, Ulrich. 1977. *Probleme der Soziolinguistik*. 2. Aufl. Tübingen: Niemeyer.

Andersen, Jack. 2005. Information Criticism: Where is it? *Progressive Librarian, no. 25*: 12-22. Retrieved February 16, 2013 from: http://web.archive.org/web/20110611100746/http://www.libr.org/pl/PL25_summer2005.pdf

Anderson, James D. & Pérez-Carballo, José. 2001a. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: research, and the nature of human indexing. *Information Processing & Management 37*(2): 231-254.

Anderson, James D. & Pérez-Carballo, José. 2001b. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: machine indexing, and the allocation of human versus machine effort. *Information Processing & Management 37*(2): 255-277.

Blake, James. 2011. Some issues in the classification of zoology. *Knowledge Organization 38*(6): 463-472.

Bliss, Henry E. 1935. *A system of bibliographical classification*. New York: H. W. Wilson.

Broadfield, A. 1946. *The philosophy of classification*. London : Grafton.

Civallero, Edgardo. 2011. UDC biology revision project: first stage: Class 59 vertebrates. http://eprints.rclis.org/16450/1/Civallero%20-%20UDC%20Biology%20Revision%20Project%20-%202011.pdf

Cooper, Rachel. 2005. *Classifying madness: a philosophical examination of the diagnostic and statistical manual of mental disorders*. Berlin: Springer.

Craig, Edward. 1998. Ontology. *Routledge encyclopedia of philosophy*, version 1.0. London: Routledge.

Ellis, David, Furner-Hines, Jonathan & Willett, Peter. 1993. Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management 3*(2): 128-149

Feinberg, Melanie. 2008. *Classification as communication: properties and design. A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy*. Washington: University of Washington. Retrieved February 16, 2013 from: http://www.ischool.utexas.edu/~feinberg/Feinberg%20dissertation.pdf

Feinberg, Melanie. 2011. *Classification as communication: properties and design*. Ann Arbor, MI: ProQuest, UMI Dissertation Publishing.

Fox, Melodie J. 2012. [Book review of two books by Rick Szostak]. *Knowledge Organization 39*(4), 300-303.

Gordon, A.D. 1987. A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A (General) 150*(2): 119-137.

Hanna, Robert. 1998. Conceptual analysis. *Routledge encyclopedia of philosophy*, version 1.0. London: Routledge.

Hansson, Joacim. 2006. Knowledge organization from an institutional point of view: implications for theoretical & practical development. *Progressive Librarian: A Journal for Critical Studies & Progressive Politics in Librarianship 27*: 31-43.

Hetherington, John. 2000. Role of theory and experimental design in multivariate analysis and mathematical modeling. In H.E.A. Tinsley & S. D. Brown eds. *Handbook of applied multivariate statistics and mathematical modeling* . San Diego: Academic Press, pp. 37-63.

Hjørland, Birger. 1992. The concept of 'subject' in information science. *Journal of Documentation 48*(2): 172-200.

Hjørland, Birger. 1998a. Information retrieval, text composition, and semantics. *Knowledge Organization 25*(1/2): 16-31.

Hjørland, Birger. 1998b. The classification of psychology. *Knowledge Organization 25*(4): 162-201.

Hjørland, Birger. 2002a. Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology 53*(4): 257-270.

Hjørland, Birger. 2002b. The methodology of constructing classification schemes. A discussion of the state-of-the-art. *Advances in Knowledge Organization 8*: 450-456.

Hjørland, Birger. 2008a. Book review of Ereshefsky (2007): The poverty of the Linnaean hierarchy: a philosophical study of biological taxonomy. *Knowledge Organization 35*(4): 255-259.

Hjørland, Birger. 2008b. Core classification theory: a reply to Szostak. *Journal of Documentation 64*(3): 333-342.

Hjørland, Birger. 2010. The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology 61*(2): 217-237.

Hjørland, Birger. 2011. The importance of theories of knowledge: indexing and information retrieval as an example. *Journal of the American Society for Information Science and Technology 62*(1): 72-77.

Hjørland, Birger. 2013a. Facet analysis: the logical approach to knowledge organization. *Information Processing & Management 49*(2): 545-557.

Hjørland, Birger. 2013b. User-based and cognitive approaches to knowledge organization: a theoretical analysis of the research literature. *Knowledge Organization 40*(1): 11-27.

Hjørland, Birger. 2013, submitted. Bibliometrics: a dynamic approach to knowledge organization. Under review in *Information Processing & Management.*

Hjørland, Birger & Hartel, Jenna. 2003. Afterword: Ontological, epistemological and sociological dimensions of domains. *Knowledge Organization, 30*(3/4): 239-245.

ISO 2788. 1986. *Documentation: guidelines for the establishment and development of monolingual thesauri*. Geneva: International Organization for Standardization.

ISO 25964-1. 2011. Information and documentation: thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva: International Organization for Standardization

Jackson, Donald A., Somers, Keith M. & Harvey, Harold H. 1989. Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist 133*(3): 436-453.

Kaplan, Abraham. 1964. *The conduct of inquiry: methodology for behavioral science*. New York: Chandler Publishing.

Mai, Jens-Erik. 2004. Classification in context: relativity, reality, and representation. *Knowledge Organization 31*(1): 39-48.

Mai, Jens-Erik. 2012. Den gode klassifikation [The good classification]. Royal School of Library and Information Science, Copenhagen, Denmark, Sept. 13, 2012. (Inaugural lecture for the professorship in information studies) . Retrieved February 16, 2013 from http://www.youtube.com/watch?v=nXLpK0JqRyM

Mammen, Jens. 2008. What is a concept? *Journal of Anthropological Psychology 19*: 25-27.

Moser, Paul K. 1998. A priori. *Routledge encyclopedia of philosophy*, Version 1.0. London: Routledge.

Moya-Anegon, Félix; Herrero-Solana, Víctor; Jimenez-Contreras, Evaristo. 2006. A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library science research and information. *Journal of Information Science, 32*(1): 63-77.

Ørom, Anders. 2003. Knowledge organization in the domain of art studies: history, transition and conceptual changes. *Knowledge Organization 30*(3/4): 128-143.

Rescher, Nicholas. 1998. Fallibilism. *Routledge encyclopedia of philosophy*, version 1.0. London: Routledge.

Rowley, Jennifer E. & Farrow, John. 2000. *Organizing knowledge: an introduction to managing access to information*. 3rd ed. Aldershot: Gower Publishing Company.

Small, Henry G. 1973. Co-citation in the relationship between two documents. *Journal of the American Society for Information Science 24*(4): 256-269.

Svenonius, Elaine. 2000. *The intellectual foundation of information organization*. Cambridge, MA: MIT Press.

Tennis, Joseph T. 2003. Two axes of domains for domain analysis. *Knowledge Organization 30*(3/4); 191-195.

Thagard, Paul. 1992. *Conceptual revolutions*. Princeton: Princeton University Press.

Toulmin, Stephen. 1999. Knowledge as shared procedures. In Yrjö Engeström, Reijo Miettinen & Raija-Leena Punamäki eds. *Perspectives on activity theory* . Cambridge, UK: Cambridge University Press. pp. 70-86.

Will, Leonard. 2008. *Glossary of terms relating to thesauri and other forms of structured vocabulary for information retrieval*. Retrieved February 16, 2013 from: http://www.willpowerinfo.co.uk/glossary.htm

Wilson, Don E. & Reeder, DeeAnn M. (eds.). 2005. *Mammal species of the world: a taxonomic and geographic reference*. 3rd ed. Baltimore: Johns Hopkins University Press.

Wilson, Patrick. 1968. *Two kinds of power: an essay on bibliographic control*. Berkeley, CA: University of California Press.

Wilson, Patrick. 1983. *Second-hand knowledge. An inquiry into cognitive authority*. Westport, Conn. : Greenwood Press.