FROM MEDIEVAL PHILOSOPHY TO THE VIRTUAL LIBRARY: A DESCRIPTIVE FRAMEWORK FOR SCIENTIFIC KNOWLEDGE AND DOCUMENTATION AS BASIS FOR DOCUMENT RETRIEVAL.

Frances Morrissey
School of Information Management & Systems
Monash University
frances.morrissey@sims.monash.edu.au

ABSTRACT

This paper examines the conceptual basis of document retrieval systems for the Virtual Library in science and technology. It does so through analysing some cognitive models for scientific knowledge, drawing on philosophy, sociology and linguistics. It is important to consider improvements in search/ retrieval functionalities for scientific documents because knowledge creation and transfer are integral to the functioning of scientific communities, and on a larger scale, science and technology are central to the knowledge economy. This paper proposes four new and innovative understandings. Firstly, it is proposed that formal scientific communication constitutes the documentation and dissemination of concepts, and that conceptualism is a useful philosophical basis for study. Second, it is proposed that the scientific document is a dyadic construct, being both the physical manifestation as an encoded medium, and also being the associated knowledge, or intangible ideation, that is carried within the document. Third, it is shown that major philosophers of science divide science into three main activities, dealing with data, derived or inferred laws, and the axioms or the paradigm. Fourth, it is demonstrated that the data, information and conceptual frameworks carried by a scientific document, as different levels of signification or semiotic systems, can each be characterised in ways assisting in search and retrieval functionalities for the Virtual Library.

Keywords

Knowledge retrieval, Virtual Library, semiotics, scientific paradigms, undiscovered public knowledge.

INTRODUCTION

This paper examines the conceptual basis of document retrieval systems for the Virtual Library in science and technology. The knowledge community that it deals with is therefore the wider scientific and technical community, rather than dealing at organisational or enterprise level as do many important knowledge management texts (e.g. Harvard 1998). Creation of objective knowledge, and its transfer and use, are integral to the functioning of scientific communities, and on a larger scale, are vital to the knowledge economy as being the core activities of science and technology.

Libraries are memory institutions for knowledge resources, and traditionally they have been repositories of published knowledge, organised for preservation, access, and use. The development of virtual or hybrid libraries is relatively recent, and is associated with the availability of digital resources and web technologies. The Virtual Library may be defined as portal or interactive gateway to one or more memory institutions of situated or distributed analogue and digital knowledge resources. Memory institutions, such as libraries, archives, and museums organise the intellectual and cultural record, including the scientific heritage, and libraries are focussed on in this instance, as dealing with published documents. Currently, virtual libraries tend to focus on traditional curatorial values for digital resources, and must develop relevant practices to support their use and management over time, including search and retrieval protocols (Dempsey 2000). However Hjorland has posited that examining problems of a cognitive nature, including the meaning of documentation, is essential for the further development of information science and of document retrieval systems (Hjorland 2000), and this type of analysis and associated outcomes are in fact the main objectives of this paper.

The principal assumption underlying the study undertaken in this paper is that libraries have been particularly important in the infrastructure of science and technology for two main reasons. Firstly, publication has long been recognised as the principal channel of formal communication in these disciplines. Second, the threefold demands in conventional scientific method of empiricism (data collection and recording), performativity (reproducibility and prediction of results), and paradigm testing, have necessitated documentation and continuing reference to collective memory in each discourse community. Indeed creating "literary inscriptions" has been said to be a central activity in experimental science, carried out at all stages from observation and instrumentation to analysis and ultimate publication (Latour 1986). However the ways in which scientists and technologists access documented knowledge, i.e., the memory traces in a discipline, bear close analysis. Hjorland (1997) has stated that the goal of information seeking in science is to identify potential knowledge, data, information, or raw material that will contribute to the theoretical or empirical development of the field or to the solution of a particular problem. Further development of the methodological aspects of information seeking, he believes, can be broken down into two aspects:

(1) problems concerning the conceptual structuring of the objects of information seeking, and
(2) problems concerning the structure and the properties of the search tools that are available to the user,
neither of which has been examined thoroughly (ibid.: 146-147). This paper attempts to redress this deficiency by addressing both the conceptual basis of scientific documents and the nature of the tools that can be used to retrieve them.

THE WORLD OF SCIENTIFIC KNOWLEDGE

The proposition that there are three types of scientific writing

Formal communication in science generally concerns information transfer through documentation and publication. Where scientists have communicated their findings by means of documents, document seeking naturally becomes the focus of attention in information seeking, and it is necessary to consider the concept of a document and the nature of a document retrieval system as metadocument (Hjorland 1997: 13-19).
It has been said that the information transferred in documents usually consists of either factual data or conceptual information, and that many studies have not made even this distinction (Meadows 1974: 92). Hjorland too has observed that information seeking tends to be routinised during normal science, concentrating on information of a factual nature, while being more to do with formulating a coherent knowledge base during theoretical transformations on the research front, thus again distinguishing only two types of information: facts and concepts (Hjorland 1997: 137-138). This paper proposes that there are in fact three types of information content, in the sense of the content of documents as embodied communicative transactions, thus differing from previous studies. This paper will categorise the three types of writing that scientists do, and thus while recognising the centrality of writing, as Latour does, it differs from Latour's view of science as an activity in which an undifferentiated spectrum of literary inscriptions are the persuasive tools in scientific argument (Latour 1986: 88). Latour's range of examples, from records of laboratory data to the published graphs and figures of formal communication in a discipline, are assumed to fit into the three categories.

Definitions of a document

In leading up to the proposition that there are three levels of meaning, first, let us consider: What is a document? In this paper, it is assumed that a scientific document, whether item (physical) or replicable event (virtual), involves an instantiation of knowledge content to observers. This is shown by considering the literal sense of "document" where "document" is derived from the Latin *documentum* (lesson, proof, instance) and still has a sense of being a teaching, or a piece of instruction or a lesson (New Shorter Oxford English Dictionary 1993 sense 1). In other words the information or knowledge content of the document has an independent existence and can be quoted, cited or challenged. The term "conceptus" is coined in this paper to denote that knowledge content. Considering possible attributes of the conceptus leads to a set of principles that could be helpful in characterising knowledge communication and access to collective memory in science.

Documents in relation to Popper's Worlds 1,2, and 3

The philosopher Karl Popper divided the universe into three "worlds". These are: World 1, the physical world; World 2, the world of conscious experiences and subjective knowledge, which mediates between World 1 and World 3; and World 3, the world of objective knowledge, including documented scientific knowledge (Popper 1996). World 3 has its own domain of autonomy. It is in part the by-product of language and of publication, particularly description and argument. "Theories, or propositions, or statements are the most important third-world linguistic entities", and they have "objective logical content." (Popper 1996: 157) Thus a document belongs both to World 1 in its physical constitution, and to World 3 regarding its knowledge component. Applying this interpretation to library and information science is not new: Swanson (1986) used World 3 as justification for his techniques for retrieving undiscovered public knowledge.

Science and the three levels of signification in scientific documents

What is science? This question is important if we are to understand the cognitive or conceptual basis of the scientific document. To answer this question, this paper will principally draw on the writings of Popper (1979) and Kuhn (1996).

Popper defines work in science as work directed towards the growth of objective knowledge (Popper 1979: 121) and as attempts to describe and (so far as possible) explain reality (ibid.: 40). He postulates that "the method of science is the method of bold conjectures and ingenious and severe attempts to refute them."(ibid.: 81). Popper describes the tasks of science as being (ibid.: 352):

(1) The derivation of predictions,
(2) technical application, and
(3) testing of the explicans, based on universal laws and specific initial conditions.

This is entirely compatible with Kuhn's listing of three foci for factual scientific investigation (Kuhn 1996):

(1) Factual determinations or increasing the accuracy and scope of factual measurements,
(2) investigations of norms and predictions: comparison of facts with predictions from the paradigm theory, and
(3) empirical work undertaken to articulate the paradigm theory, such as
    (a) determination of universal constants,
    (b) formulation of quantitative laws, and
    (c) elucidation of models

where the paradigm is the overarching conceptual framework accepted by the discourse community.
Even Einstein, known as a physicist rather than as a philosopher of science, recognised three main themes in his schematic of science (Einstein 1952):

(1) The system of axioms (which I identify with the paradigm),
(2) deduced laws, and
(3) the totality of sense experiences (with which I include the outputs of instrumentation)

I would like to propose that these three foci of science correspond with three important possible facets of scientific documents, whether experimental or theoretical: the reported or implied data, the surmised information or laws and trends, and the overarching conceptual framework or paradigmatic level of scientific documentation.
Popper comments "All the important things we can say about an act of knowledge consist of pointing out the third-world objects of the act - a theory or proposition - and its relation to other third-world objects, such as the arguments bearing on the problem as well as the objects known" (Popper 1979: 163). This makes it easy for us to equate Kuhn's paradigms with Popper's World 3 objects.

The problem of induction and the phony story, contrasted with deduction

Popper makes a useful distinction between the "bucket theory of science" and the "searchlight theory of science". In the former, "...Our knowledge, our experience, consists either of accumulated perceptions (naïve empiricism) or else of assimilated, sorted and classified perceptions..." That is, the mind resembles a container in which perceptions and knowledge accumulate, from which objective experience or science may be distilled (Popper 1979: 341). In this model, the purpose of scientific documentation is to construct a phoney narrative in which scientific hypotheses pose as inductions or as analogical reasoning. This style of writing is deplored by Latour (1986: 174), because it leads to a separation of actual recorded laboratory practice and factual statements in the propounding of novel insights, and to the stripping of contingent circumstances (ibid.: 174), i.e., "scientific facts are formulated in the denial and obliteration of their own historicity" (ibid.: 277). In the alternative, the searchlight theory of science, particular hypotheses inform what kind of observations to make, and which observations deserve attention; and observations also function as tests of hypotheses. It is worth noting that Berkenkotter and Huckin have also noted the construction of the "phoney story" in scientific reporting, and they observe that as part of this construction, scientific genres are expected to demonstrate both novelty and intertextuality with regard to previous expositions of the paradigm (Berkenkotter & Huckin 1995).
Thus scientific documents may be structured in such a way as to either have the appearance of inductive reasoning, with concomitant citation of others' results and conclusions as well as one's own, or to be the result of deduction from the axioms, models and metaphors making up the paradigm. In either case, there are at least two levels of signification present, either explicitly or by implication.

Hjorland on seeking knowledge objects

The contemporary information scientist Hjorland (1997: 147) lists four categories of concepts and tools that can be used in categorising and seeking knowledge objects:

(1) the research object itself,
(2) scientific theories and methods concerning the research object,
(3) scientific disciplines, traditions, and paradigms that deal with the research object,
(4) formal aspects of the communication process where knowledge about the research object is documented

These correspond with the threesome of empiricism, rules and predictions, and the paradigm, with the documented knowledge *per se* in its documentary genre making up Hjorland's fourth category.

The threefold path

In this section I have drawn on the arguments of major philosophers of science to conclude that science constitutes the pursuit of objective knowledge, with such knowledge being reported and transferred in the form of scientific documents. These philosophers seem unanimous in their opinion that science and scientific inscriptions are made up of three main threads, viz.: data, being original observations and outputs of instrumentation; trends, rules and predictions drawn from either phenomenological generalisations or theoretical derivations; and the paradigm. I extrapolate this conclusion to declare that any scientific document whether experimental or theoretical will carry or imply meaning associated with two or more of these threads.

CONNECTIONS WITH CONCEPTUALISM, FROM OCKHAM TO PRICE

Conceptualism

William of Ockham (ca. 1285-ca. 1347) and Henry Habberley Price (a mid-20th century philosopher) have much to say on the nature of concepts that is useful to understanding information and knowledge retrieval. Taken in conjunction with the afore discussed nature of scientific knowledge, their views lead to possible new functionalities for document retrieval interfaces.

What are concepts? First take the modern view (Audi 1999: 169): Concepts are principles of classification empowering recognition, analysis and categorisation, and they may be divided into 4 classes. These are mental representations (ideas serving classificatory purposes); brain states that serve the same function; general words and their usage; and the ability to classify correctly using one or more of these. Henry Price was the main proponent of this view (ibid.) Second, take Ockham's view: sciences, together with their premisses and objects, are collections of true propositions (Kretzman et al 1982: 505), where Ockham believed that mental propositions exist, and are composed of understandings or concepts (Ockham 1991: 206-209). Ockham stipulated that a concept is a natural sign of a thing, either in actuality or in a *de possibili* proposition(ibid.: 454-456).

These views are important and insightful for the Virtual Library, because from the 17th century onward, hence entirely within the era of print-based technologies, nominalism and phenomenalism had overtaken conceptualism. That is, it was believed that experience and sense-data were the foundations of observation and generalisable truths, and that there were no universals or transcendentals that had independent existence. This runs directly counter to Popper's World 3, described in the previous section. It is, however, the basis for much 20th century philosophy, such as that of Willard Van Orman Quine. Quine built a philosophy of science based on linguistics and formal symbolic logic, which is understandable in terms of the sense of closure associated with print and its definitive embedding of the oral, spoken word in visual space (Ong 1982). Quine's view is that "words and observable behaviours are all we have to go on" (Quine 1995: 6) and "sentences are the primary vehicles of meaning" (ibid.: 7). Science in his view is built on sensory data and scientific posits, the latter being in the form of occasion sentences, standing sentences, and observation categoricals, and subject to truth and verdict functions (Quine 1974: 3-80, 1995: 43). This view, although laudable, does not seem particularly helpful in interface design for the Virtual Library. The Virtual Library and Web-based technologies have multimedia capabilities which extend beyond the semantics of written and spoken language. These capabilities can be compared with views that originated in a preliterate society, which are that concepts have an independent existence, and that these are important in the transmission of culture, including scientific culture. Exploring these views in this section is built on assumptions that information seeking is primarily to do with identifying, accessing and retrieving documents that embody or exemplify particular concepts, be they to do with subject, authorship or other criteria.

"Intentions" as signs signifying concepts

Ockham divided concepts into "first intentions" and "second intentions". A first intention is a natural sign of something which is not itself a sign. A second intention is a natural sign that signifies other natural signs or first intentions (Audi 1999: 492).

Concepts and first and second intentions lead to an ability to categorise, as follows. A category has three sorts of existence: in the mind, in writing, and in speech. A category, taken significatively, is composed of concepts(Ockham 1991: 475 - 478). If a category is taken as an ordering of things then a category may include both first intentions and second intentions. If category is taken as what is first and most common in a categorial ordering then each category is a first intention or name of first intention (Ibid.: 467 - 471). Also, in Ockham's view, concepts may be distinguished as being absolute, connotative or relative(ibid.: 485-487). This has obvious relevance to modern classification schemes, which are commonly descriptive, denotative and/ or hierarchical. For example, take a recent author, Gail Hodge, on the details of knowledge organisation systems for digital libraries.

Hodge (2000) lists 3 characteristics common to knowledge organisation systems:

(1) The knowledge organisation system imposes a particular view of the world on a collection and the items in it (i.e. it operates at paradigm level).

(2) The same entity can be characterised in different ways depending on the knowledge organisation system used.

(3) There must be sufficient degree of resemblance between the concept expressed in a knowledge organisation system and the real-world object or representation it applies to, for others to apprehend and apply the system with reasonable reliability.

She lists 3 main types of knowledge organisation system:

(1) Term lists,

(2) Classifications and categories, and

(3) Relationship lists.

As extensions of these, she gives examples of ways of linking digital library resources to related resources, and of indexing with and mapping of multiple schemes. Hodges' 3 main types of KOS seem to correspond directly with Ockham's division of concepts into the connotative, the absolute and the relative.

Applications of insights from conceptualism to knowledge representation

The twentieth century philosopher Price said, "The occurrent manifestation of concepts is as multiform as the exercise of intelligence is. In fact it is just another name for the exercise of intelligence." (Price 1953: 354). He went on to explain that intelligence is manifested in thinking, or conceptual cognition, which enables the recognition of objects, classification, .and the formulation of complex ideas through intellectual operations performed upon the basic concepts. Concepts are simply recognitional capacities founded in the similitude of things, or awareness of universals (ibid.: 357).

Price believed that "fundamentally a concept is a recognitional capacity, whatever else it may be besides." (Price 1953: 355). He stated that concepts are also manifested through:

(1) sign cognition, or understanding signs through mental and psychophysical response. This includes both primary recognition and secondary recognition (where the latter can be mistaken), and induction involving the operators *not*, *or* and *if*, as well as statistical probability, which allows for probabilities or degrees of instantiation. (In my view this has to do with empiricism in science, and also the nature of declarative knowledge. It correlates with the first level of scientific activity after Popper and Kuhn, described in the previous section.)

(2) production of quasi-instantiative particulars such as production and recognition of images (i.e. "image thinking", both quasi-instantiative and generic), production of replicas, and production of instantiative particulars (instantiation in action in real life). (In my view this has to do with both reproducibility and hypothesis testing in science. It correlates with the second level of scientific activity after Popper and Kuhn, described in the previous section.)

(3) production of non-instantiative symbols (words and codes, and symbolic operations), and the ability to produce alternative verbal formulations (again, degrees of actualisation). (In my view this has to do with modelling, mathematical formalism, and formal communication in science. It correlates with the third level of scientific activity after Popper and Kuhn, described in the previous section.)

Price stressed the need for cognition of signs and symbols (Price 1953: 352-353):

> The noise or mark, so long as I merely observe other people producing it, is not yet a symbol for me, but at most only a sign. It only becomes a symbol for me when I have learnt to use it for myself. After all, every man must understand for himself what others say or write. The others cannot understand for him. It is his concepts, not theirs, which are operative when he 'follows' what his neighbours say. His concepts, and not only theirs, must be associatively linked with the sounds or marks which he hears or sees. And that is only another way of saying that he must already be able to use these words understandingly in his own thinking. Each of us must understand for himself, and no one else can understand for him.

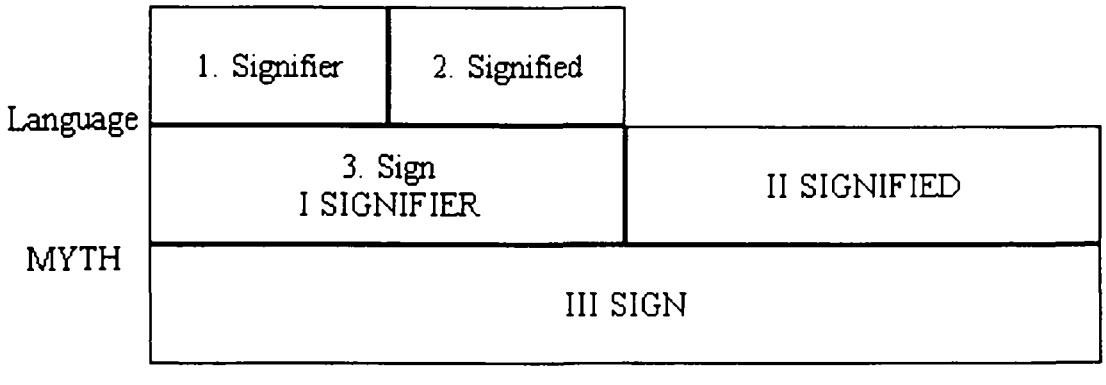
In other words, knowledge retrieval must deal not only with the text at data level (the marks or words) but the concepts as well. Both must be capable of being recognised, and this has much to do with the science of semiotics, which will be discussed in the next section. This is where it will be seen in detail how much medieval scholasticism and modern conceptualism have to offer modern information science. Also, conceptualism and the three levels of analysis of scientific knowledge that have been discussed (data and intentions, information such as trends, rules and predictions, and conceptus or paradigm), will be shown to feed into a new model of scientific documentation with potential retrieval points for documents, in Section 5.

HIGHER ORDER SEMIOTICS

Roland Barthes is well known for his writing on semiotics and myth (Barthes 1972). His ideas will be discussed in this section as an attempt to lay more epistemological and ontological foundations for considering document retrieval systems, answering the question "What is knowledge?" with the response "A level of meaning in a higher order semiotic system". I shall draw parallels between the theoretical and referential content of scientific communication and Barthes' "myth", later to be compared with Kuhn's "paradigm".

Barthes regards mythology as a branch of semiotics, where semiotics is concerned with signification, and analyses communication in terms of the functional relationships between the triad of signifier, signified and sign. To quote Barthes (1972: 114-115; emphasis in original).,

> ...Myth is a peculiar system, in that it is constructed from a semiological chain which existed before it: it *is a second-order semiological system.* That which is a sign (namely a concept and an image) in the first system, becomes a mere signifier in the second...
>
> It can be seen that in myth there are two semiological systems: a linguistic system, the language (or the modes of representation which are assimilated to it), which I shall call the *language-object,* because it is the language which myth gets hold of in order to build its own system; and myth itself, which I shall call *meta-language,* because it is a second language *in which* one speaks about the first.



(Barthes 1972: 115)

Figure 1: Myth as second order semiotic system

I see the signification of scientific documents as usually involving wider or deeper meaning and inferences, signified beyond the conventional language or formulaic representations of the information- or data-content of the document as a physical item. That is, there is more than the basic signification of the words and symbols of the text. Trends and paradigms may be inferred from mere data by a knowledgeable reader; and data may be predicted from axioms and laws given particular contexts and boundary conditions. There is therefore more than one level of signification, and at each level, the triadic relation of signifier-signified-sign applies. The semiotic approach to the levels of signification is described in detail in the following section.

DOCUMENT RETRIEVAL FUNCTIONALITIES BASED ON CHARACTERISATIONS OF DOCUMENTED DATA, INFORMATION AND CONCEPTUS

The following section expounds a novel analytical approach to scientific documents and documentation, based on the philosophical discussions of the previous sections. The analysis distinguishes between data, information and knowledge as different levels of meaning, and likens them to levels of semiotic systems. Based on this analysis, characteristics of the documented data, information and knowledge are enumerated. It is proposed that these could be used as descriptors, identifiers or allocated or intrinsic metadata for document access and retrieval.

Signification at data level and associated retrieval functionalities

I postulate that data consists of primary signs and significations (words, numbers and other meanings), and use of language. It includes "first intentions" and "second intentions", and representation of acts of understanding, such as the empirical content of primary scientific literature, and of informal communications. It corresponds with the first of Kuhn's three foci for scientific activity.
Data can be identified and retrieved through specification and characterisation of the item or its data content in allocated or intrinsic metadata (where the item is the physical document, i.e. the medium and encoding existing to transfer information in the process model of communication).
In semiotic terms, the triad of signifier, signified and sign at the data level may be represented in the diagram below.

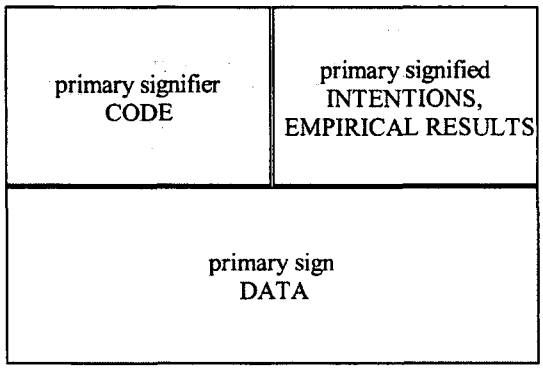


Figure 2: The semiotic triad for data

Signification at information level and associated retrieval functionalities

I postulate that information is the meaning drawn from or associated with data or selected data, such as trends, rules and predictions. It corresponds with the second of Kuhn's three foci for scientific activity. It is carried in physical

and virtual texts and formal and informal scientific communications, and often said to be the content of the process model of communication. The field, tenor and mode of communicated information in a discourse community are conventionalised as aspects of genre (Schirato and Yell 1996).
Information is amenable to classification through use of subject analysis. Thus retrieval of information is possible through specification and characterisation of the item or its information in allocated or intrinsic metadata. These forms of metadata include bibliographic citations, metadata associated with recognised metadata schemes,and subject representation in classification schemes, indexing, or surrogate records (whether by literary warrant or user warrant). In most libraries, the indexer/ cataloguer assigned bibliographic authority files (subject, author, added entries) are important. Information can also be identified and retrieved through its lexical content as item full text.
In semiotic terms, the triad of signification at information level may be represented in the diagram below. It can be seen to overlay the data triad, in that data has moved from being the sign to the signifier.

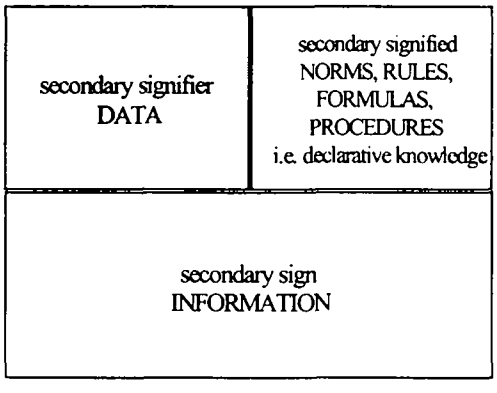


Figure 3 The semiotic triad for information

Signification at conceptus level and associated retrieval functionalities

I postulate that the "conceptus" or atomisation of overarching conceptual framework consists of knowledge, theory, and their essential contexts. These include the referential content of scientific genres, in the third of Kuhn's three foci for scientific activity. The conceptus may demonstrate novelty, intertextuality and promulgation of paradigm in "normal science", or convey competing articulations and explicit discontent with the current paradigm during "extraordinary science" (see Kuhn 1996, and Berkenkotter & Huckin 1995). The conceptus may be disseminated as a "work" or "teaching" (see the earlier discussion on derivation of "document").
The conceptus may be identified and retrieved through identifiers, through subject analysis of the paradigm (by analogy with the case of information), and through citation indexing. The aspects of imprimatur, or intellectual authority and history of the conceptus, consist of other details such as citation trails, corporate authorship, sponsorship, and publishing filters such as refereeing. These details, which acknowledge where credit is due for current or past work, may also be denoted by any details of intellectual property transfer, reformatting or reinvention which may arise, but which are not covered above.
The manifest identity and uniqueness of the conceptus are associated with its ideation, as we found in discussing conceptualism, and therefore the mental states associated with a concept are as important as its lexical content. In present times, with multimedia available on the WorldWideWeb, the ideation of the conceptus can be represented in

the author's, or demonstrated by the user's, sets of hypertext relations for the item. These hypertext relations may be hierarchical, associative or chaining.
In semiotic terms, the triad for conceptus may be represented in the diagram below. It can be seen to overlay the triads for data and information.

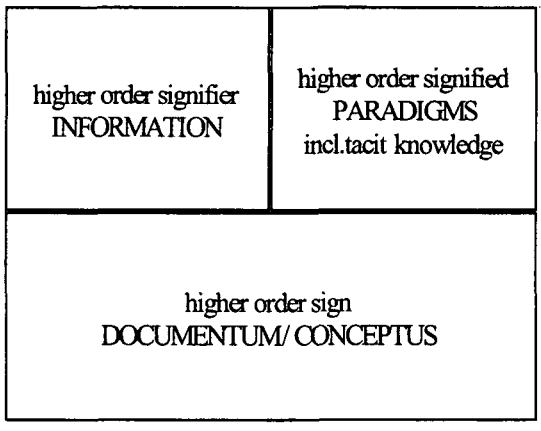


Figure 4 The semiotic triad for knowledge and conceptus

Undiscovered public knowledge

Undiscovered public knowledge is identified and retrieved through literature-based discovery from syntaxis of any or all of the concepts listed above, i.e. from conjoint retrieval of item metadata and/or any other descriptors and identifiers of the documented information or oeuvre. The retrieval of undiscovered public knowledge is based on the presumption that Popper's World 3 contains implicit knowledge that has not to date been stated explicitly; and contains increasing amounts of undiscovered public knowledge as the combinations and permutations of concepts exceed the number of concepts (Swanson 1986, Davies 1989).
Davies (1989) has identified a number of categories of undiscovered public knowledge:

(1) hidden refutations or qualifications of hypotheses;
(2) inferences from transitive relations (partial syllogisms or inference chaining, found through cocitation analysis);
(3) cumulative weak tests (large sets of documents with weak evidence);
(4) unrecognised or hidden analogies;
(5) hidden correlations.

Based on previous analysis in this paper, Davies' numbers (1), (2) and (4) operate at paradigm level; (3) operates at data or information level; and (5) operates at information level.
Swanson has extensively tested one particular approach for finding undiscovered public knowledge. He has made it available as the "Arrowsmith" front-end for processing citations from Medline, at http://kiwi.uchicago.edu/index.html. The approach uses data mining on title keywords, and is described by Swanson and Smalheiser (1997). In brief, the methodology is statistical analysis of title keywords from "complementary literatures", assuming

(1) the existence of an intermediate concept B (which may be conjecture initially) between a concept represented in a set A of articles and a concept represented in a second set C of articles
(2) that this concept can be deduced from putative co-citation of keywords B in sets A and C
(3) and that hypotheses may be generated and/or substantiated through examination of documents retrieved using these keywords.

In terms of conceptualism, the method operates on the lexical content of the item or metadata; and in terms of Davies' list, it belongs to the category of inference from transitive relations, where keywords are assumed to act as markers of partial syllogisms AB and BC, inferring AC.

This digression is intended to demonstrate that the descriptive framework for scientific knowledge being presented in this paper is compatible with other studies, and, importantly, has practical applications.

As another (and light-hearted) aside, some Virtual Library interface functionalities making use of the attributes of documents, through understanding the levels of signification, can be remembered in the MAPROOM + X mnemonic:

MP: Metadata Parser,
ROO: Repositories of Organised Objects (i.e. physical or virtual collections using knowledge organisation systems),
M: MetaCrawler for discovery in Web-based full text resources,
X: literature-based discovery through semantic and syntactic characterisation of item and conceptus

Synopsis of insights from semiotics

The hierarchical nature of the levels of description of documented knowledge can be represented as a tree-like structure, to show figuratively that information arises from data and its relations; and that scientific knowledge at paradigm level is anchored in information and its relations. Alternatively, the levels of scientific meaning or signification can be illustrated in semiotic terms. The discussion in this paper to date and the models presented of data, information and knowledge are summarised in the following diagram, which clearly shows the overlaying of the semiotic triads.

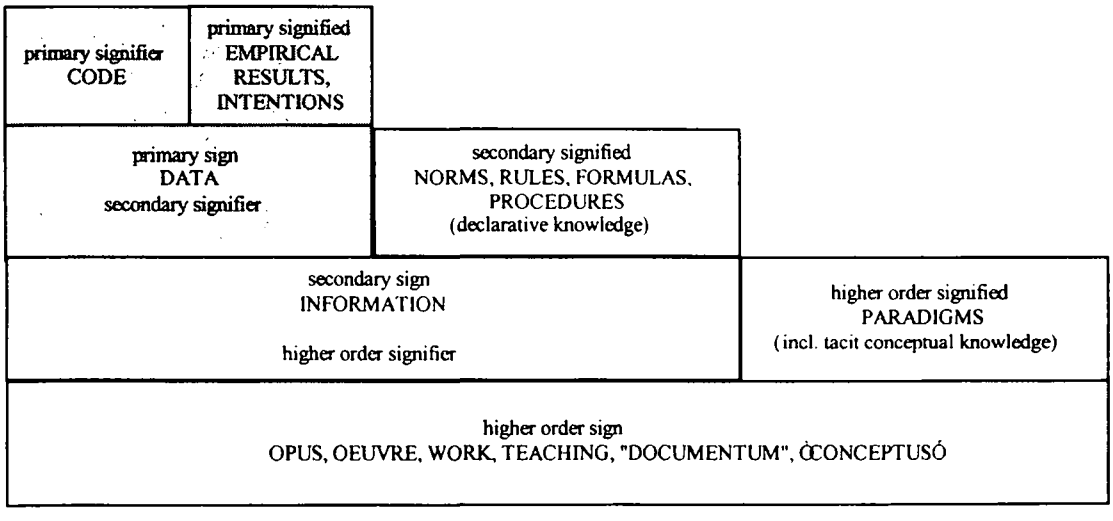


Figure 5 The overlapping semiotic triads for data, information and conceptus

The insights summarised above, lead to a dyadic construct for a scientific document as two entities: the physical item (the medium plus encoded data or information as first and second order semiotic systems respectively); and the metaphysical "conceptus" (the signified ideational content or knowledge component, at a third or higher order of

semiotics, which in science is associated with paradigms or conceptual frameworks, as per Kuhn 1996). Postulating recursion of this type is not new. Tursman (1987: 46) states that "a sign is potentially, if not in fact, a member of an infinite series of signs". He quotes Peirce's 1895 writings (Eisele 1976 IV: 309):

> A sign stands *for* something *to* the idea which it produces, or modifies. Or, it is a vehicle conveying into the mind something from without. That for which it stands is called its *Object;* that which it conveys, its *Meaning;* and the idea to which it gives rise, its *Interpretant.* The object of a representation can be nothing but a representation of which the first representation is the interpretant... The meaning of a representation can be nothing but a representation... So there is an infinite regression... Finally, the interpretant is nothing but another representation, to which the torch of truth is handed along; and, as representation, it has its interpretant again. [Lo] another infinite series.

As foreshadowed in an earlier section of this paper, documented scientific knowledge may be characterised at any of the three levels of semiotic system. For example, Einstein's theory of special relativity may be characterised at conceptus level, with credit apportioned in citation trails, leading back to the original publication; or in scientific textbooks it may be treated in terms of the mathematical equations derived from it (information level, with application of laws and formularies); or it may be investigated at empirical level, with published, "grey" or informal reports of data collected in experimental studies. And in a second example, weather observations and meteorology can be dealt with at the level of data, seasonal and climatic trends, and their probable causation.

I would like to conclude this section by stating that there is the potential to build on and develop further some of the attributes of documented knowledge as semiotic systems, in order to increase the effectiveness of Virtual Library search and retrieval interfaces. The concepts associated with the ideation of the conceptus, and its aetiology and imprimatur, are especially ripe for investigation.

CONCLUSION

It has been demonstrated in this paper that conceptualism and philosophy of science from the medieval scholar William of Ockham through to Thomas Kuhn lead to a dyadic construct for a scientific document as two entities, one physical and one metaphysical: the item (the medium plus encoded data or information as first and second order semiotic systems); and the "conceptus" (the signified ideational content or knowledge component, which in science is associated with paradigms or conceptual frameworks). It is proposed that the item at data and information level, and the conceptus as separate entities, have attributes enabling their effective characterisation and management. Further, it is demonstrated that any scientific document by virtue of its subject content will carry meaning significant at three levels: data; laws, trends and predictions; and the overlying set of axioms or paradigm. These attributes can be regarded as principles informing knowledge communication and access to collective memory in science. This paper suggests that the principles just described could form the basis for a number of distinct types of functionality or functional metadata for a Virtual Library interface. To utilise these functionalities, metadata characterising the conceptus may have to be introduced or used more widely. The insights obtained in this paper, and the potential search/ retrieval functionalities they lead to, could increase the efficiency and effectiveness of knowledge access and transfer in scientific communities, and by extension, help reduce lead times for innovation in the knowledge economy.

Acknowledgements

While undertaking the research covered in this paper, the author has been in receipt of a scholarship funded by the School of Information Management & Systems, Monash University; and by a SMURF grant to Dr Frada Burstein, SIMS (Knowledge Management for Information Communities). And an earlier version of this paper was presented at the ACKMIDS (Australian Conference for Knowledge Management and Intelligent System Support) 2000 conference 4-5 December 2000, Melbourne, Australia.

REFERENCES

Audi, R. (Ed.) (1999) **The Cambridge dictionary of philosophy**. 2nd ed. Cambridge: Cambridge University Press.

Barthes, R. (1972) **Mythologies**. trans. by A. Lavers. New York: Hill and Wang.

Berkenkotter, C. & Huckin, T.N. (1995) **Genre knowledge in disciplinary communication: cognition/ culture/ power**. Hillsdale NJ: Lawrence Erlbaum.

Brown, L. (Ed.) (1993) **The new shorter Oxford English dictionary.** Oxford: Clarendon Press.

Davies, R. (1989) The creation of new knowledge by information retrieval and classification. **Journal of Documentation** 45 (4) December 1989 pp. 273-301.

Dempsey, L. (2000) Scientific, industrial, and cultural heritage: a shared approach; a research framework for digital libraries, museums and archives. **Ariadne** http://www.ariadne.ac.uk/issue22/dempsey.html (accessed 29 February 2000).

Einstein, A. (1952) Letter of 7 May 1952 to Maurice Solovine, cited by Miller, A.I. (1984) **Imagery in scientific thought.** Cambridge MA: MIT Press p. 45.

Eisele, E. (1976) **The new elements of mathematics by Charles S. Peirce vols. 1-4.** The Hague: Moulton, quoted by Tursman (1987: 46).

Harvard business review on knowledge management. (1998) Boston: Harvard Business School Press.

Hjorland, B. (1997) **Information seeking and subject representation: an activity-theoretical approach to information science.** Westport CT: Greenwood Press.

Hjorland, B, (2000) Documents, memory institutions and information science. **Journal of Documentation** vol. 56 no. 1 January 2000 pp. 27-41.

Hodge, G. (2000) **Systems of knowledge organization for digital libraries: beyond traditional authority files.** Washington DC: Digital Library Federation, Council on Library and Information Resources.

Kretzman, R., Kenny, A., & Pinborg, J. (Eds.) (1982) **The Cambridge history of later medieval philosophy: From the rediscovery of Aristotle to the disintegration of scholasticism 1100 – 1600.** Cambridge: Cambridge University Press.

Kuhn, T.S. (1996) **The structure of scientific revolutions.** 3rd ed Chicago: University of Chicago Press. (First published 1962.)

Latour, B. & Woolgar, S. (1986) **Laboratory life: the construction of scientific facts.** Princeton NJ: Princeton University Press.

Meadows, A.J. (1974) **Communication in science.** London: Butterworths.

Ockham, W. of (1991). **Quodlibetal questions, volumes 1 & 2. Quodlibets 1-7.** Translated by Freddoso, A.J. and Kelley, F.E. Newhaven: Yale University Press.

Ong, W.J. (1982) **Orality and literacy: the technologizing of the word.** London: Routledge.

Popper, K.R. (1979) **Objective knowledge: an evolutionary approach.** Rev. ed. Oxford: Clarendon Press. (First published 1972.)

Price, H.H. (1953) **Thinking and experience.** London: Hutchinson.

Quine, W.V. (1974) **The roots of reference: the Paul Carus lectures.** La Salle, IL: Open Court.

Quine, W.V. (1995) **From stimulus to science.** Cambridge MA: Harvard University Press.

Schirato, T. & Yell, S. (1996) **Communication and cultural literacy: an introduction.** Sydney: Allen & Unwin.

Swanson, D.R. (1986) Undiscovered public knowledge. Library Quarterly vol 56, April 1986, pp. 103-118.

Swanson, D.R. & Smalheiser, N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. **Artificial Intelligence** vol 91, 1997, pp. 183-203. Available http://kiwi.uchicago.edu/webwork/Alabtext.html (accessed 20 July 1999).

Tursman, R. (1987) **Peirce's theory of scientific discovery: a system of logic conceived as semiotic.** Bloomington IN: Indiana University Press.