The Unified Medical Language System: an epistemological perspective

Hana Storova

Masters in Library and Information Science Thesis
Royal School of Library and Information Science, Copenhagen, Denmark
Acknowledgments

The author would like to express special thanks to her supervisor Birger Hjørland for his guidelines, help and fruitful discussions during the course of this research.

The author would also like to thank to her family and friends for their support along the way.
Abstract
This thesis analyzes the Unified Medical Language System, a unique project of biomedical vocabulary integration developed by the National Library of Medicine, U.S., and evaluates it from the perspective of theories of concepts implied by the main epistemological positions. The view of meaning represented in the metathesaurus, the main component of the system, is analyzed and compared with the representation of meaning provided by the source vocabularies. Finally, the theoretical principles upon which the UMLS is developed are analyzed and system is evaluated from the epistemological perspective, especially from the perspective of theories of concepts related to historicism and pragmatism.

Keywords:
Unified Medical Language System, epistemologies, knowledge organization, concepts, semantics
Table of Contents

1 Introduction ........................................................................................................................................... 6
2 Unified Medical Language System ......................................................................................................... 8
   2.1 Project overview, history and origin ................................................................................................. 8
   2.2 Metathesaurus ................................................................................................................................... 10
      Metathesaurus concept entry ............................................................................................................. 13
   2.3 Semantic network ............................................................................................................................. 15
   2.4 Lexicon ............................................................................................................................................ 17
   2.5 Representation of the meaning in the UMLS .................................................................................... 17
      Semantic structure of the metathesaurus ............................................................................................ 19
   2.6 Problems of vocabulary integration .................................................................................................. 20
   2.7 Current research and use of UMLS .................................................................................................. 22
3 Metatheories and epistemologies in information science ...................................................................... 24
   3.1 Information science and philosophy ................................................................................................. 24
   3.2 Metatheories and paradigms in information science ......................................................................... 25
   3.3 Theories of knowledge and their importance ................................................................................... 26
      Empiricism ........................................................................................................................................ 27
      Rationalism ....................................................................................................................................... 28
      Positivism ......................................................................................................................................... 28
      Logical positivism ............................................................................................................................... 29
      Historicism ....................................................................................................................................... 29
      Pragmatism ....................................................................................................................................... 30
      Importance of theories of knowledge ................................................................................................. 30
   3.4 Epistemologies in information science ............................................................................................. 31
   3.5 Knowledge organization and theories of knowledge ......................................................................... 32
   3.6 Paradigms in medicine ..................................................................................................................... 36
   3.7 Implications for knowledge organization ......................................................................................... 38
4 Concept theory and semantics in knowledge organization ................................................................... 39
   4.1 Importance of semantic issues ......................................................................................................... 39
   4.2 Research on concepts ....................................................................................................................... 39
   4.3 Concepts and semantic issues in knowledge organization ............................................................... 42
   4.4 Theories of knowledge and their view of concepts ......................................................................... 44
5 Understanding the meaning in the UMLS ............................................................................................. 48
   5.1 Biomedical vocabulary integration .................................................................................................. 48
      Multiple epistemological perspectives on the UMLS ..................................................................... 49
      Multiple meaning representation ........................................................................................................ 49
      Epistemological perspective ............................................................................................................... 52
6 Conclusion ............................................................................................................................................. 57
7 References ............................................................................................................................................... 61
8 Appendix ............................................................................................................................................... 65
   8.1 Semantic Types ............................................................................................................................... 65
   8.2 Semantic relations ............................................................................................................................. 67
   8.3 Abbreviations ................................................................................................................................... 67
   8.4 Figures ............................................................................................................................................. 67
“Language is the source of misunderstandings”

(Antoine de Sait-Exupery)
1 Introduction

Assisting the user in finding the relevant information is one of the most challenging tasks to be solved in information science (IS). Two disciplines of IS that are mostly concerned with finding the solution to this problem are information retrieval (IR) and knowledge organization (KO). One of the primary goals of KO is the creation of various semantic tools, such as thesauri, semantic networks and controlled vocabularies that aid information retrieval and knowledge organization purposes. The problem of navigating and accessing information resources is especially important for scientists in various fields, including the field of biomedicine which is the subject of interest in this thesis. The more diverse the research field is, the more diverse are its language and information resources. Various knowledge organizing systems (KOS) are constructed to connect the language of users and the language of documents and assist in IR. The problem of finding relevant information is further complicated by the use of different semantic tools such as controlled vocabularies in various databases and information resources of the field.

The key research focus of this thesis, a long-term research project of the National Library of Medicine called the Unified Medical Language System (UMLS, http://www.nlm.nih.gov/research/umls/) is an attempt to find a solution to these problems, caused by the diversity of information sources as well as the diversity of terminologies and controlled vocabularies that have been used in the biomedical domain. The UMLS joins various KOS in biomedicine into one common system and enables translations of terms from one vocabulary to another. Its three main components are metathesaurus, semantic network and SPECIALIST lexicon.

This thesis tries to analyze theoretical principles upon which the UMLS is developed and evaluate the system from epistemological perspective. Theories of knowledge, such as empiricism, rationalism, historicism and pragmatism, underlie various approaches to knowledge organization and they imply different methodological ideals of how knowledge organizing systems should be constructed. The principles of KOS construction should relate to theories of concepts and semantics because KOS, including UMLS, are in fact semantic tools representing concepts and semantic relationships. The main epistemologies imply also a specific view of concepts. The view that has been termed “post-Kuhnian” and that is related to historicism and pragmatism regards concepts as developed in relation to various paradigms and scientific theories that are present in every domain, including biomedicine. Concepts are
seen as embedded in social and historical contexts. This view has specific implications for the principles of KOS construction. These issues will be considered when evaluating the UMLS as a semantic tool that represents biomedical concepts and meaning and serves the purposes of retrieval and integration of biomedical information.

The discussion of these issues is structured in the following way:

Chapter 2 describes the UMLS, its history and origin, and the development of its main components, metathesaurus and semantic network. Semantic structure of the metathesaurus, in which biomedical concepts play the key role, is described and the main principles of the metathesaurus construction are discussed. Brief overview of problems related to the integration of biomedical vocabularies into the UMLS follows. Finally, the intended use of the UMLS and its current applications and research are described.

The role of the philosophy in information science, most importantly in knowledge organization, is addressed in chapter 3. It is argued that deep theoretical assumptions, that may be called metatheoretical or epistemological, are at the core of all theories and approaches in our field. The main theories of knowledge (empiricism, rationalism, positivism, historicism, pragmatism) and their characteristic features are described. The role of epistemology in various theories in KO is emphasized. It is demonstrated that theories of knowledge have fundamental impact on how knowledge organizing systems are constructed.

The importance of semantic issues in IS and KO is addressed in chapter 4. Because KO is concerned with the development and construction of KOS which are semantic tools, it should relate to the theories of concepts and semantics. Brief introduction to main theories of concepts follows. Different views of concepts and semantics characteristic to the main epistemological positions are introduced.

Finally, chapter 5 analyzes the view of meaning provided in the UMLS and discusses its relation the meaning represented in the source vocabularies. UMLS is evaluated from the perspective of the main theories of knowledge, especially from pragmatic perspectives. It is argued that UMLS disregards the importance of social context and historical development of concepts, as understood in the post-Kuhnian view. However, the UMLS recognizes and transparently represents the differences between meaning in various source vocabularies because they are functional and they reflect the diversity of purposes for which the vocabularies were developed.

Chapter 6 concludes that although UMLS does not fulfill the requirements of a fruitful KOS from the
historicist and pragmatic perspective, it tries to find the compromise between the context-relativeness of meaning and the need for biomedical vocabulary integration which is necessary for integration and retrieval of diverse biomedical information.

2 Unified Medical Language System

2.1 Project overview, history and origin

The Unified Medical Language System (UMLS) has been developed and directed by the National Library of Medicine (NLM), U.S. since 1986. This long-term research project is a result of cooperation of many groups of experts from NLM and many other research and academic institutions, including the American Medical Association, and some private agencies and professional associations. The project may be considered as a front research in terminology and knowledge organization.

The main purpose of the UMLS is to facilitate retrieval and integration of biomedical information from diverse information resources [Humphreys, 1989] and to improve the conceptual connections between the language of users and various machine-readable information [Humphreys, 1993]. The initial idea was to create a set of semantic tools and related programs that could be implemented in various information systems which provide access to diverse biomedical information to a large community of health researchers and practitioners. These smart interfaces would use the UMLS semantic tools to map user queries to diverse biomedical information and to identify the most useful information sources. In this way, the UMLS would enable the computer systems to “understand” biomedical language.

The UMLS is intended to solve the problem of finding biomedical information that is spread in various sources and that is moreover difficult to access because of the diversity of terminologies used in these sources. It is designed to compensate for the differences in various terminologies and coding schemes used in biomedicine. Linking these terminologies together would make it possible to provide an efficient and integrated access to diverse biomedical information resources, such as bibliographic and factual databases, expert systems, computer-based patient records and other resources.

The main components of the UMLS are advanced semantic tools called “Knowledge Sources” that are used by smart interface applications and programs. The three semantic tools defined in the beginning of the project were: (1) a metathesaurus that would contain terms from various biomedical vocabularies and would also contain contextual information from these source vocabularies together with representing inter-term relationships and semantic type of each term, (2) semantic network which would be a separate structure that would define broad semantic categories of the terms (called semantic
types) and valid relationships among these categories, and (3) information sources map which would provide information about the biomedical information resources themselves, i.e. about their scope, vocabulary, syntax, location and access conditions [Humpreys, 1989]. Later, another tool called SPECIALIST Lexicon was created for managing the linguistic variation in natural language and in the source terminologies themselves.

It was also decided that in the metathesaurus the individual source vocabularies would be integrated on the concept level, using a central representation of each concept and mapping each vocabulary only once, to this central canonical form.

The first phase of the UMLS project focused mostly on assessing needs and defining the main components. The first experimental versions of the UMLS metathesaurus and semantic network were released in 1990 and the first version of information sources map a year later. The development of a range of applications and smart programs that could use the knowledge sources followed. Whereas the knowledge sources were developed mainly centrally, with NLM playing the key role in the process, the applications that make use of them were developed in a decentralized manner, and the system has been tested and improved in many real life applications within different areas of health care research and practice.

Since the start of the project, the UMLS has been gradually improved through a range of different versions, updates and approximations. During the development process, the versions of knowledge sources were broadly disseminated and improvements and changes in the system were guided by user feedback, testing and evaluation. The UMLS metathesaurus gained a lot in its scope and coverage. The UMLS Knowledge Source Server (umlsks.nlm.nih.gov/) which allows users and developers to navigate and retrieve information from data files was made available on the Internet. Recently, the latest UMLS 2009 AA release was made available through the new improved version of the UMLS Knowledge Source Server.

It has been emphasized many times that the UMLS is not an attempt to develop a single standard biomedical vocabulary or single biomedical knowledge base and that UMLS is designed to compensate for the differences in the various biomedical vocabularies while recognizing their importance [Humpreys, 1989; Lindberg, 1993; Tuttle, 1988]. From the early beginning of the project, the UMLS metathesaurus has been envisioned as a thesaurus that transcends the individual source vocabularies from which it is built and that makes the differences in medical terminology transparent to the user. In
the initial phases of the project the creation of a new canonical classification of biomedical concepts, to which existing vocabularies could be mapped, was considered [Tuttle, 1988; Lindberg, 1993]. This approach was, however, found too demanding and unnecessary to achieve the goal of the UMLS to aid retrieval and integration of biomedical information from various sources. Instead it was decided that the various terms and alternative concepts names from the individual source vocabularies will be directly linked in the metathesaurus.

Thus, the statement that the UMLS is not a single universal biomedical terminology is related to the fact it was not built as a new comprehensive terminology but from the existing biomedical vocabularies and to the objective of the UMLS that tries to preserve the original meaning and contexts of its source vocabularies and represent the differences between them. Its intended purpose is to link the various terminologies on the level of concepts, to enable mappings and translations among them and to aid retrieval of biomedical information from a vast range of sources.

These are the reasons why some authors argue that the main value of the UMLS is that it provides a common framework for the individual vocabularies and that it should not be evaluated as a terminology itself because that puts the UMLS in a competing position with its source vocabularies [Campbell et al., 1998a].

The UMLS metathesaurus is nevertheless an advanced semantic tool consisting of a huge number of biomedical terms, concepts and a rich network of both semantic and lexical links. Its scope and complexity is bigger and its granularity finer than any of its source vocabularies. It is therefore only natural that it may be evaluated as a knowledge organizing system. Its potential role and value resides in its ability to represent a large part of biomedical terminology and to serve as a reference tool other vocabularies may use, as it already case with many terminologies, e.g. Medical Subject Headings (MeSH) or Gene Ontology. Biomedical vocabularies often use the information from the UMLS to map their terms to other KOS and to resolve various terminological issues.

2.2 Metathesaurus

The most important part of the UMLS is the metathesaurus, a large database containing information about biomedical concepts, their various names and relationships among them. The metathesaurus also provides definitions of the concepts and other value-added information. Biomedical concepts which constitute the content of the metathesaurus are derived from a large variety of terminologies, classifications and controlled vocabularies developed for various purposes in different areas of
biomedicine and health care.

An important feature of the metathesaurus is the transparent representation of synonymy. A particular entity (disease, treatment, diagnostic method etc.) may be represented in different terminologies by various names, although it refers to the same meaning. Biomedical terms that are considered as synonyms are in the metathesaurus put together to form a single concept. Concepts in the UMLS are clusters of synonymous terms. Thus, the various biomedical vocabularies are integrated on the concept level, whereas their various hierarchical contexts and other information is preserved. The original hierarchical information is provided for every source vocabulary, even if it disagrees with the metathesaurus view.

To explain the integration of terms in more detail: A term from a particular source vocabulary is first mapped to the metathesaurus concept. Using automatic lexical methods and subject specialist revision, the term is assigned to the matching metathesaurus concept. In this way, many different terms from existing vocabularies that are considered synonyms are put together to form a concept.

However, each concept entry represents different views of the semantic neighborhood of the concept. The metathesaurus preserves the different views of its source vocabularies. It is possible to see what position a particular term had in the original structure, what parent, child, siblings and other related terms it had in different vocabularies. In the metathesaurus, these differences can be viewed in the concept entry. This contextual information is represented together with the UMLS own view of semantic relations.

The metathesaurus is thus organized by meaning and its basic unit is a concept, represented by a single metathesaurus entry. The main purpose of the metathesaurus is to link synonymous terms and alternative concept names and to identify useful semantic relationships among concepts [National Library of Medicine, 2008].

Concepts in the metathesaurus are linked to other concepts through a small number of broadly defined relationships [McCray, 2002]. These relationships are either inherited from the structure of the source vocabularies or they are introduced by other methods during the metathesaurus construction. In any case the relationships are reviewed by the UMLS editors. These include basic kinds of semantic relationships, i.e. hierarchical, associative and equivalence relationships. Also statistical relationships (so called co-occurrence information) derived mostly from the co-occurrence of MeSH indexing terms in MEDLINE citations are available.
All concepts are moreover assigned into broad semantic categories, so called semantic types, which define the meaning category of a particular concept (semantic type is for example disease, syndrome etc.). Semantic categories are described and defined in the second UMLS Knowledge Source called semantic network. Each concept is assigned to one or more semantic types from the semantic network.

The metathesaurus has been described as a thesaurus that transcends the individual vocabularies because it provides a detailed map of both semantic and lexical links between the terms and because it encompasses the scope of its source vocabularies and organizes it by meaning. Its scope is bigger and its granularity finer than any of its source vocabularies [Lindberg, 1993; Schuyler, 1993].

The metathesaurus contains terms from more than 100 source vocabularies of many different types. These include large disease and procedure classifications used for statistical reporting and billing purposes, terminologies designed for the use in patient-record systems; and also more specialized vocabularies (used in psychiatry, genetics, etc.) and disease terminologies used in the expert diagnostic systems; thesauri used in information retrieval and many others. The major categories of the source vocabularies include: diagnosis (25%), procedures and supplies (25%), diseases (19%), comprehensive vocabularies (15%), and others (e.g. anatomy, drugs, genetics, nursing, miscellaneous). Some vocabularies belong to more categories.

All the source vocabularies included in the metathesaurus were originally in English and still represent the major part of the system (62%). Recently also vocabularies in other languages such as Spanish, French, Dutch, Italian, Japanese, Portuguese and Czech have been added.

The scope of the metathesaurus is determined by the scope of the source vocabularies from which the concepts and their names and relationships are taken, although many of the concept attributes and relationships and even some concepts are also added by the NLM during the process of metathesaurus development. In theory even a concept that is not present in any source vocabulary might appear in the UMLS metathesaurus.

Each concept entry in the metathesaurus represents one single meaning, i.e. one particular concept. Alternate names for the same concept, such as synonyms, lexical variants and translations are all connected together in the concept entry. First, the strings that are lexical variants of each other are grouped together to form a single term. Terms that are considered as having the same meaning are than linked together as alternate names of the same concept [Lindberg, 1993]. All the terms and concepts names belonging to the same concept, that is with the same or closely related meaning, are linked
together by a special code called Concept Unique Identifier (CUI). One of the terms that belong to the concept is identified as a preferred term. Preferred terms are taken from the list of source vocabularies ranked according to a certain priority. Preferred terms are taken from the source with highest priority. This function of the metathesaurus can, however, can be adjusted in the individual applications that implement UMLS.

Concepts, lexical variants (concepts names), strings (graphical and spelling variations of a concept name) and atoms (occurrences of a particular string) may be considered as the basic units from which the metathesaurus is constructed. Beside concepts, lexical variants, strings and atoms are also assigned and linked through unique identifiers.

**Metathesaurus concept entry**

Concept information available in the metathesaurus for the concept “Hypertensive disease” includes:

**Concept name**: Hypertensive disease

**Concept Unique Identifier (CUI)**: C0020538

**Definitions**: definitions of the concept from five source vocabularies.

**Semantic Type**: Disease or Syndrome

**Synonyms/Lexical variants**: different terms and expressions used in various source vocabularies to name the concept “Hypertensive disease”. For example:

*Blood Pressure, High; Hypertension; Hypertensive disease; Hypertension, arterial; Blood Pressure, Increased; Hypertensive vascular degeneration* etc.

**Relations**: children, parents, broader, similar, narrower, other related, related and possibly synonymous, siblings, source asserted synonymy concepts are represented; the source vocabulary of this semantic information is always indicated.

**Contexts**: ancestors, children and siblings of the particular term (belonging to the concept class in the metathesaurus) represented in the particular source vocabulary (e.g. MeSH)

**Co-occurrence information**: provides a view of terms that co-occurred with the concept as two main indexing terms, mostly in MEDLINE citation records
Metathesaurus search view in the UMLS Knowledge Source Server is represented on Figure 1.

2.3 Semantic network

Whereas the metathesaurus represents biomedical terminology from a large number of different vocabularies organized by meaning, the second UMLS Knowledge Source called semantic network provides a structure that encompasses and unifies all the source vocabularies. It defines broad semantic categories called semantic types, to which every metathesaurus concept is mapped, and it also defines the relationships between the semantic types. Semantic network represents the knowledge about the biomedical domain. It has been described as “upper-level ontology” of that domain [McCray, 2003].

There are 135 semantic types in the semantic network and each concept is assigned at least one, and usually not more than five semantic types. The function of the semantic types and relationships defined in the network is to help with the interpretation of the meaning of every concept. Terms in the metathesaurus are linked at the individual concept level, but the structures and hierarchies of individual source vocabularies are to a certain extent preserved. However, by the means of the higher level
semantic categories (semantic types) and their relationships, concepts can be uniformly interpreted at
the more general level of the semantic network, regardless of the structure of the original source
vocabulary. The semantic network structure thus allows a consistent view of all metathesaurus concepts
[McCray, 1995].

Semantic type entry in the semantic network contains a definition of the semantic type, information
about position of the semantic type in the hierarchy of semantic network, i.e. information about its
parent(s) and children, and information about what semantic relationships can exist between the
semantic type and other semantic types. As an example, the semantic relationships plausible between
the semantic types 'Antibiotic' and 'Disease or Syndrome' are:

\[
\text{Antibiotic affects / causes / complicates / diagnoses / prevents / treats Disease or Syndrome}
\]

Semantic types are divided into two broad categories: entity and event (see appendix for details).
Semantic types exist in the different levels of specificity and granularity. Always the most specific
semantic type available is assigned to a concept in the metathesaurus. Some examples from the list of
current semantic types are: physical object, organism, plant, substance, idea or concept, finding,
language for the entity category; and activity, technique, biologic function, cell function, injury or
poisoning for the event category.

There are fifty-four semantic relationships defined by the semantic network (see appendix for details).
The most common and basic one is the “isa” relationship, which links most semantic types. The
hierarchy of semantic types is established through the “isa” semantic relationship. The example of 'isa'
semantic relationships is: Human “isa” Mammal. There are also non-hierarchical associative
relationships in the UMLS Semantic Network, main of which are: \textit{Physically related to, Spatially
related to, Temporally related to, Functionally related to, and Conceptually related to}. These additional
relationships distinguish the network from the simple type hierarchy.

Each semantic relationship is also associated with a definition, and a set of semantic types that can be
plausibly linked by this relationship. The level of granularity and specificity of relationships varies
across the semantic network. It is important to note that the semantic relationships defined by the
semantic network provide meaning links between two particular semantic types, but they do not
directly link two concepts. The relationship defined between two semantic types is a possible meaning
link between two broad meaning categories but it may or may not hold for any particular pair of
concepts [McCray, 1995].
2.4 Lexicon

The third UMLS Knowledge Source called SPECIALIST Lexicon (here referred to as lexicon) contains morphosyntactic information. This large lexicon was developed to provide the lexical information needed for the SPECIALIST Natural Language Processing (NLP) System. The lexicon contains both commonly used English words and biomedical vocabulary from a variety of sources, such as the UMLS test collection of MEDLINE abstracts, Dorland's Illustrated Medical Dictionary, The American Heritage Word Frequency Book, Longman's Dictionary of Contemporary English, current MEDLINE citation records and others and the UMLS metathesaurus. Each lexicon entry provides syntactic, morphological (inflection, derivation and composition information) and orthographic (spelling) information about a particular term, needed for SPECIALIST NLP System.

In the UMLS, the lexicon and related lexical tools are used for managing the linguistic variation in the biomedical terminologies as well as in natural language. They are used to suggest related terms in the process of terminology integration in the metathesaurus. The lexicon and lexical programs may be also used independently of other UMLS knowledge sources, in various NLP applications [McCray, 2002].

2.5 Representation of the meaning in the UMLS

The UMLS metathesaurus is organized by meaning. Its important feature is that it links synonyms, i.e. the terms and concept names from source vocabularies with the same or closely related meaning, together. The synonymous terms that together form a concept class are either inherited from the source vocabularies (they are regarded as synonyms in one of the source vocabularies), or they are discovered by lexical matching and lexical resemblance methods. In any case, what view of synonymy to represent in the metathesaurus concept structure decide the UMLS editors which are subject experts in biomedical domain.

Not all terms belonging to the same metathesaurus concept have strictly the same meaning. Synonymy are defined more loosely and sometimes even closely related terms are considered as synonyms. The reasons are practical. Clustering of more or less synonymous terms into concepts is the main principle of the metathesaurus through which the goal of linking the different biomedical vocabularies and enabling translations and mappings between them is achieved [Bodenreider, 2001].

While the synonymy and other semantic relationships taken from the source vocabularies are in the UMLS adjusted and the resulting UMLS view is often different from the view of the source vocabularies, it is the aim of the UMLS to preserve the original meaning and contexts. Thus, the
synonymy and other semantic relationships from the source vocabularies are also represented, even if they do not agree with the metathesaurus view. This means that when for example the same term is used to refer to two or more different concepts in some of the source vocabularies, the metathesaurus represents all the meanings and also indicates the source vocabulary of each meaning. In another situation, when the same concept is placed in different positions within the structure of various source vocabularies (i.e. has different ancestors, siblings and children), the metathesaurus represents all this semantic information. Even when there are conflicting relationships between two concepts in different vocabularies, they are also both included in the metathesaurus.

Therefore, UMLS metathesaurus is not a single, comprehensive classification/ontology of biomedical domain. Only on the higher level of semantic types assigned to the individual concepts, there is a certain unified view of concepts and semantic relationships in biomedical domain.

Hierarchical relationships are important in providing the structures in the semantic tools such as thesauri and classifications. Traditional medical classifications are monohierarchical, i.e. they have a single tree architecture in which a concept can be identified by its position in the tree hierarchy. Because UMLS takes some of the hierarchical relationships from its source vocabularies which structure differs, it allows for multiple inheritance creating a graph structure.

There are two different kinds of hierarchical relationships in the UMLS metathesaurus. By convention, the relationships called parent/child relationships come from the source vocabularies. Although they have usually been defined in a source vocabulary on the term-level, in the metathesaurus they represent relationships between concepts. There is also another kind of hierarchical relationships called broader/narrower relationships that are added to the structure using different methods, such as automatically by lexical analysis of terms or they can be established by human editors [Bodenreider, 2001]. These two types of hierarchical relationships may be sometimes redundant. However, the use of lexical methods enables to discover relationships among concepts coming from different source vocabularies that cannot be inherited from the structure of any individual vocabulary.

Two major kinds of hierarchical relationships are hyponymy or generic relation that is represented by the 'isa' relation or by 'narrower than' relation. Another kind is meronymy represented by the 'part of' relation.

Another type of relationships represented in the UMLS structure are mapping relationships, i.e. relationships that point to terms in another vocabulary (external mapping). Most of these relationships
(around 90%) are taken from the Medical Subject Headings in which so called supplementary concepts are mapped to Medical Subject Main Headings. They are considered as hierarchical relationships in the UMLS since the supplementary concepts usually provide finer granularity than the main headings.

**Semantic structure of the metathesaurus**

The semantic structure of the UMLS is a result of joining three different perspectives: (1) semantic information derived from the source vocabularies, (2) the metathesaurus own view, (3) the structure provided by the semantic network. These three perspectives together give create a rich network of biomedical concepts, their names and variety of semantic and lexical links, which in fact provides a model of a biomedical domain.

The metathesaurus is organized by the principle of semantic locality; concepts with closely related meaning are linked together. Each concept has its particular semantic neighborhood which is made of the concepts semantically related to a particular concept, both in formal and informal way.

Concepts in the metathesaurus can be seen as related based on four different aspects of semantic locality [Nelson, 1991]:

1) **Term information:**

   Every concept in the metathesaurus is identified by a class of synonyms. Terms can be synonyms represented in a concept entry either because some of the source vocabularies list them as synonyms, or because they are suggested as synonyms by lexical matching techniques. In both cases the synonyms are reviewed by human editors.

2) **Source contextual information:**

   Relatedness of terms can be also inferred from their position in the structure of the particular source vocabulary, i.e. from source contextual information. Each source vocabulary provides a different view of the semantic neighborhood of a concept (different set of associations between the concept and related concepts). Thus, the metathesaurus in fact represent a multiple views of semantic neighborhood of a concept.

3) **Semantic Type:**

   Concepts are also related by sharing the same semantic type. All the concepts assigned to the same broad semantic category are linked together.

4) **Data Co-occurrence:**

   18
Terms may be alternatively viewed as related based on co-occurrence of these terms. This method is used to empirically discover that terms are related because they co-occurred together, e.g. as two major index terms in the same article. For example co-occurrence of two MeSH terms in the same article in MEDLINE citations records implies a relationship between these terms.

Nelson, et al. [1992] found that the degree of overlap of these four different dimensions of semantic locality is relatively small and that each of them provides a unique representation of the semantic neighborhood of a concept.

2.6 Problems of vocabulary integration

The UMLS metathesaurus integrates biomedical terminology from many different vocabularies. The process of integration of a new vocabulary into the UMLS has several steps [Lomax, 2004]. First, the overlap between the coverage of the new vocabulary and the UMLS is determined. The purpose and structure and of the vocabulary are studied by the UMLS team of experts with the help of the source vocabulary developers. The vocabulary is algorithmically assigned semantic types from the UMLS semantic network and the assignment is reviewed by human editors. Next, Specialist lexical programs and other heuristics are used to automatically map the terms from the vocabulary into existing concepts in the UMLS. Each mapped concept is reviewed by the editors, and in some cases also modified.

During the integration process many terminological issues must be solved. For some concepts there may not be a corresponding semantic type available in the UMLS semantic network. Quite often the source vocabularies have finer granularity of semantic categories than the semantic network. It might thus be the case that two concepts that were assigned two distinct semantic categories in the source vocabulary will be assigned the same semantic type in the UMLS. However, the policy of the UMLS is to assign a concept to the most suitable semantic category available, and rather than adding new types into the semantic network, the broader semantic types is assigned.

The main problem in integration of synonyms into the metathesaurus is that synonymous relationships are context-relative [Bodenreider, 2001]. They may hold true in one source vocabulary that has a certain granularity and in which it is functional to establish synonymous relationships between particular terms. They may however become invalid in the context of another vocabulary with different granularity and with generally different functionality. Not all terms regarded as synonyms in the source vocabularies are thus incorporated as synonyms in the metathesaurus. Sometimes they are split into two
different concepts. In other cases, however, even terms that are only quasi-synonyms are incorporated as synonyms into the same concept entry.

In general, implicit contextual information present in the source vocabularies constitutes another problem in the integration of terms into the metathesaurus. Not all the terms in the biomedical vocabularies are syntactically correct and fully specified entities and they are therefore hard to discover by lexical resemblance techniques. This is especially the case with terminologies that were not made for computational purposes and where some term information or part of the term expression may be omitted because it is assumed to be clear from the context and stays therefore implicit. In situations like this, the context of terms must be restored by human editors before incorporating them into the metathesaurus.

Synonyms may also change over time, they may split into two or more different concepts. Also a terms or expressions may gain several different meanings. On the other terms having two different meanings may be merged into one concept. The metathesaurus keeps track of merges, splits and deletions in its source vocabularies and makes updates accordingly.

Hierarchical relationships are often implicit in the structure of some source vocabularies and sometimes generic and partitive relations are mixed together. This causes a lot of problems in the integration of relationships into the metathesaurus. The relationship „location of“ which is considered as non-hierarchical in the metathesaurus is considered as hierarchical in some source vocabulary. As a result, the relationship is integrated as hierarchical from its source and at the same time represented as non-hierarchical in the metathesaurus structure. For example a disease may be connected through the parent/child relationships with the part of the body in which it is located although it is not right from the ontological perspective. Sometimes the integration of hierarchical relationships results in circular hierarchical relationships, in which a concept becomes both ancestor and descendant of another concept, or even of itself (loop relationship) [Bodenreider, 2001].

Differences in granularity of individual source vocabularies also cause problems in integration of hierarchical relationships into the metathesaurus. Whereas in one vocabulary with lower granularity two terms may be considered as siblings, i.e. they are on the same level in the hierarchy, in a vocabulary with finer granularity they may be hierarchically related.
Also the presence of not fully specified terms in some source vocabularies resulted in invalid relationships in the integration. In particular inter-concept relationships discovered by lexical techniques suffered from the problems related to implicit and contextual knowledge.

Sometimes the ambiguity of terms in the source vocabulary is resolved implicitly, by placing the terms in the different place within the structure of the vocabulary. These terms are than found ambiguous during the integration, as was the case when mapping the Gene Ontology (GO) into the UMLS. For example, GO uses the same terms for referring to enzymes and also the activity of a particular enzyme. The ambiguity is implicitly resolved by that enzyme terms were children of 'molecular function' in the vocabulary hierarchy. However, during algorithmic mapping of GO terms into UMLS, these two meanings were put together as they shared identical text strings. This problem is that of multiple concepts with the same concept name and different meanings and one that has to be resolved during the integration of a new vocabulary into UMLS. As a result, the naming of a GO concepts referring to enzyme functions was changed into “enzyme activity” both for the purpose of mapping and in the GO vocabulary itself.

This is an example of how the process of mapping a vocabulary into the UMLS helped to resolve problems with term ambiguity in a source vocabulary. The problem of the same string with different meanings is quite common in GO, especially where the same term applies to different species, but with a different meaning in each. These ambiguities gradually evolved in the language over and can be only resolved in their context of use [Lomax, 2004]. Another problem in integrating GO was described as the need to differentiate between the true synonyms and terms broader, narrower or just related to the term, which were in GO all considered as synonyms but during integration had to be considered as separate concepts.

The example of mapping GO into the UMLS demonstrates the fact that the source vocabularies from which the metathesaurus derives semantic information were developed for various purposes and this has impact on the way they classify terms. This has also impact on their integration into the UMLS.

**2.7 Current research and use of UMLS**

The UMLS Knowledge Sources were not originally developed for the use by the end-users, i.e. health care practitioners and researchers themselves. They were designed to be used by information system developers who would implement them in biomedical information systems that help users to find relevant information from various sources. Thus in most cases the Knowledge Sources would stay
invisible for end-users.

Thus, the situation for which the UMLS is tailored is the following. Health researcher or practitioner is looking for information on a particular subject or topic. An interface program interprets his search question using the information from the UMLS metathesaurus and semantic network. It identifies important concepts and their relationships in the query. With the help of the information from metathesaurus or other Knowledge Sources, the interface program helps the user to clarify and re-formulate the query. Then the interface program matches the identified concepts, semantic types and relationships with a particular database or information source, described in the information sources map. The interface program selects the potentially useful databases, run the searches in them and than merge, rank and display the results to the user [Humphreys, 1993]. This is the ideal scenario the UMLS project is trying to achieve. Clearly, both advanced Knowledge Sources and sophisticated interface programs play essential role in the success of this scenario.

However, because the Knowledge Sources provide a rich network of biomedical concepts, their various names and relationships, and they integrate the individual thesauri into a common framework while preserving the contextual information from these vocabularies, they may also be useful as reference tools for librarians.

The structure of the metathesaurus should enable the users to perform various tasks, such as collecting terms used to name a particular concept, extracting the relations of a particular concept to other concepts, or obtaining a set of concepts for a broad semantic category-semantic type [Bodenreider, 2004].

National Library of Medicine and other institutions use the UMLS Knowledge Sources in a range of applications, including information retrieval, natural language processing, creation of patient and research data, and the development of enterprise-wide vocabulary services.


Research and Quality [National Library of Medicine, 2006].

During more than two decades of the project development, large number of research related to different areas of UMLS development, maintenance, use and application was conducted and published. The National Library of Medicine bibliography about UMLS from the years 1986 to 1996 provides 280 citations [Selden and Humphreys, 1997]. These include papers discussing the foundation of the project; development of individual Knowledge Sources and their structural and semantic properties; UMLS applications in areas such as natural language processing, indexing and retrieval, or vocabulary construction and concept discovery; and various preliminary studies related to the project. Journal articles from PubMed/MEDLINE covering articles published since 1997 to the present provided by NLM contain 882 citations.

3 Metatheories and epistemologies in information science

The following chapter discusses the role of philosophy, especially theories of knowledge, in information science. It is argued that epistemological assumptions are behind the various theories in our field. Main theories of knowledge also imply specific views of how knowledge organization should be done and how knowledge organizing systems should be constructed.

3.1 Information science and philosophy

Before the development of modern science, scientific questions were subject of “natural philosophy” which was part of metaphysics. Philosophy played important role in science until the 18th century when the individual sciences became independent of philosophy. The 20th century was strongly influenced by logical positivism which regarded philosophy as unimportant and only later, with the emergence of new alternative philosophical approaches such as hermeneutics, pragmatism or critical and realist philosophies, the importance of philosophy of science could be reconsidered. Recently, there has been growing interest in the philosophy of science in many special fields, including for example medicine or economics.

In library and information science, as well as in many other special fields and disciplines, the role of the philosophy of science has been considered as limited or that of rather small importance [Hjørland, 2005b]. However, various philosophical questions are present in many issues in information science and to avoid them means to keep the discipline non-academic and practice-oriented, without a chance for advancing the field and creating its own theories.
The role of philosophy is to develop more general understanding and to solve theoretical issues that should help the individual special sciences. If information science wants to develop a more general knowledge about information systems, information retrieval, subject analysis and representation and other issues, that could be useful for other special disciplines, it must relate to philosophical and theoretical issues. It should work more closely with certain aspects of the philosophy of science, most importantly those that are related to epistemologies.

Information science must formulate the theoretical assumptions and principles that are behind the different approaches and practices within the field Hjørland [1998, 2005b]. Without trying to achieve this, information science can never gain respect from other special sciences.

3.2 Metatheories and paradigms in information science

Every field has its theories that provide a set of assumptions, principles and explanations of certain part of the reality the particular field studies. Theory is a tool for explaining and understanding some phenomena. A theory in IS may, for example, explain information seeking behavior of users or provide an explanation of how different kinds of subject representation contribute to the efficient information retrieval or how information systems should be built to meet certain needs and requirements. In IS, theories from other fields such as psychology or sociology are often used and not many theories are theories explicitly developed within the field itself [Hjørland, 1998]. Also, theories in IS often only describe and predict how a certain phenomenon could work, but they rarely provide a real explanation of that phenomenon. They are thus still at what Bates [2005] calls “modeling stage” of forming a theory and have not really become theories in the right sense. However, specific approaches in IS, such as information retrieval or subject representation, have deeper theoretical basis, certain assumptions that have been called “metatheoretical” [Hjørland, 1998].

An important concept in the philosophy of science is a metatheory as a philosophy behind a theory or theories. According to Bates [2005, p. 2], metatheory is “the philosophy behind the theory, the fundamental set of ideas about how phenomena of interest in a particular field should be thought about and researched”. Metatheories have been also termed “paradigms”, “research traditions” or “schools”. All these terms refer to the theories that describe, investigate, analyze and criticize the theories in a domain [Hjørland, 2005b]. Metatheoretical assumptions are broader than theories, are connected to philosophical views, and are behind theoretical, empirical and practical work in every domain. Metatheories or paradigms are often interdisciplinary theoretical trends and can be identified across the
domains.

Metatheory also usually implies a certain methodology of a field. Metatheory is in fact a combination of the philosophical and theoretical understanding and critique of a particular theory and the methodology resulting from this understanding. Thus, theories are often implicitly connected to a metatheory and corresponding methodology [Bates, 2005].

Whereas in the natural sciences one paradigm is usually dominant and the emergence of a new paradigm means fundamental change of the field, in the social sciences there are often more metatheories or philosophies of research co-existing and competing with each other. This is also the case in the library and information science.

Bates [2005] provides an overview of 13 metatheories that currently exist in information science. As it is typical for social sciences, many different metatheories exist in information science next to each other and influence various areas of research in the field. Each metatheory may have a few different interpretations and researchers may move from one to another metatheory in their work. Bates divides metatheories in two main groups. Those that aim at establishing general laws and theories in their investigations, as it is typical in the natural sciences, are called *nomothetic*. Those that concentrate on description of single and unique facts and particulars and derive understanding but not general rules from these descriptions are called *idiographic*. This approach is mostly characteristic for the humanities. Both nomothetic and idiographic approaches are important and necessary for the understanding and progress in a field. Social sciences are according to Bates influenced by both approaches. Different metatheories in information science may thus be more in correspondence with one or the other or combination of both. To the first approach according to Bates belong historical, constructivist, constructionist/discourse-analytic, philosophical-analytic and critical theory approach. To the second belong cognitive, bibliometric, physical, engineering, user-centered design and evolutionary approach. On the intersection are ethnographic and socio-cognitive/domain-analytic approaches.
3.3 *Theories of knowledge and their importance*

Epistemology is a philosophical discipline that studies knowledge. It is concerned with what knowledge is and also with what ways of obtaining knowledge are valid. Epistemological assumptions are at the core of theories in information science and they have implications for how these theories view the objects of their interest, how they study and evaluate them. These assumptions may be more or less implicit, but they always play a very important role in a particular theory of IS.

In every field including information science, it is possible to identify various metatheories and paradigms, which are the deep theoretical trends behind the theories, as well as various theories of knowledge, which are even more general theories about what knowledge is and how it can be obtained and thus also about how research and science should be done.

The value of epistemologies is in their potential to serve as a deep theoretical framework in which each theory or approach in IS (or another field) can be analyzed and compared with other theories, and its relative strengths and weaknesses can be discovered. Epistemology does not provide final or best solution to the problems of the field. It can, however, help to uncover theoretical assumptions behind different approaches in the field. In this way the problems inherited in these approaches can be discussed and solved, and the field can make progress.

Hjørland [1998, 2003, 2005a, 2005b, 2008b] identifies and discusses main theories of knowledge and analyzes how epistemologies influence different theories and approaches in information science. The main families of epistemologies are empiricism, rationalism, positivism, historicism and pragmatism. It is important to discuss and consider these theories of knowledge because they underlie theoretical and practical issues within the different approaches in information science. The following section provides a brief overview of the main epistemologies [see e.g. Craig, 1998; Hollis, 1994; Markie, 2008]

**Empiricism**

Two classical epistemological positions that have been the most influential in the history of science, especially in the European thinking in the period between scholasticism and Kant, are empiricism and rationalism. The main difference between these two approaches lies in what they perceive as a primary source of knowledge. Empiricism traditionally regarded as a primary source of knowledge human sensory experiences and “objective” observations of these experiences. Human being is according to empiricist view born without any knowledge, and all the knowledge a person obtained through senses. The more complex knowledge is regarded as a result of synthesis of more simple, primitive pieces of
knowledge that came from senses. Thus, even the more complex knowledge came, though indirectly, from experience. In practical research, empiricism implies collecting of simple, verified observations on which any observer can agree on (so called intersubjectivity) and from which a more general observations can be induced. Thus, empiricism is a bottom-up approach whose two basic methods are observation and induction. Relevant methods of obtaining information are in empiricist view observations and sensory data and inductions from the collections of this data. Regarded as irrelevant are data about observer's assumptions, point of view or pre-understanding.

The main empiricist assumption to which other more historicist and interpretative epistemologies object is that experiences are “given”, and that experiences and the knowledge we obtain through them are independent of our theories and conceptualizations, and of the influence of culture and society.

**Rationalism**

In contrast to empiricism, rationalism is a philosophy that favours reasoning and a priori theorizing to experience and observation. At least in its traditional form, rationalism regards sensory experience as inferior and unimportant. A central theses of rationalism is that at least some of our concepts are innate and do not come from experience. Such a priori, inborn knowledge is seen as a precondition for any experience. From rationalist perspective concepts are inborn structures which classify our perceptions. Rationalism prefers deductive methods to inductive. Relevant methods of gaining information according to rationalism are a priori theorizing and logic, mathematical and computer models, and systems of axioms, definitions and theorems. Empirical data are considered as less important and inferior.
Although traditionally put in opposition, empiricism and rationalism have one important feature in common – they both believe that knowledge can and should be “objective” and the only question is how to obtain this neutral, value-free, “given” knowledge that represents or mirrors neutral, objective reality. Although empiricism perceives as a valid way to obtain knowledge observations and experiences whereas rationalism believes in theorizing and a priori knowledge, both these positions believe that knowledge should be a mirror of objective natural order. Both rationalism and empiricism share the view that knowledge has objective basis and that the methods of obtaining knowledge are independent of the researcher's perspective and theories. They believe that objective natural order in the universe can be discovered by human knowledge in a way that is not relative to specific point of view. This view has been criticized by other alternative epistemologies, most of which can be considered as belonging to a family of epistemological positions related to historicism (see below).

**Positivism**

Positivism is one of the most influential philosophies in the history of Western thought. It is often associated with empiricism for its rejection of metaphysics and for its views that science must be empirically and experimentally based. The term was coined by August Comte, one of the first modern sociologists, in the middle of the 19th century. Positivism emerged from different movements in the philosophy and social science in the late 19th century and even Comte's ideas differ from the way positivism is generally understood. While positivism as understood by Comte admitted a central role of theory for science, researchers that are influenced by positivism often refuse the fact that their own views and investigations are influenced by theoretical and cultural issues.

Positivism rejected metaphysics and believed in one, universal a priori scientific method and in that the role of philosophy was to study this method and also become 'scientific'. Positivism also believed that all sciences may be reduced to physics and that in this way science can be united. A theory in a good science must in positivist view be translatable into observational statements and terms. In the social sciences positivism emphasized quantitative data and formalized theories and regarded the study of meaning and interpretations as scientifically inadequate [Kinciad, 1998].

**Logical positivism**

Closely related to empiricism and rationalism is a philosophy called logical positivism. It shares with empiricism the view that all the knowledge comes from observation and senses. Logical positivism combines empiricism and rationalism in that it joined the empiricist emphasis on observation with
rationalist logic. Logical positivism was also skeptical of metaphysics. This position implied the division of science in two parts, empirical and observational part on one hand, and theoretical and formal part on the other hand. Logical positivism studied mostly linguistic and logical aspects of knowledge. Like positivism, logical positivism believes in the unity of sciences and in one universal scientific method. It also holds the view that sciences form a hierarchy and that physics is the most fundamental discipline into which all other sciences may be reduced.

**Historicism**

Historicism regards knowledge as socially and culturally determined. According to historicism, perception, observation are influenced by researcher's pre-understanding and theories, and knowledge is culturally and socially embedded. Historicism stands in opposition to positivism (as well as empiricism and rationalism) in that it regards context, background and culture as important for understanding any human activity, including creation of knowledge, science or philosophy. Cultural and historical contexts serve as a frame of reference in historicist philosophies. Historicism considers knowledge about historical developments, contexts, pre-understanding, theories and conceptions as relevant. Interpretation is the key method in historicism, whereas empirical data without context which cannot be interpreted are considered of limited importance.

Historicism is often associated with famous philosopher Thomas Kuhn [1962] whose work “The structure of scientific revolutions” explains the progress of science from historical perspective. According to Kuhn, science makes progress with so called “paradigm shifts” in which one scientific paradigm replaces another paradigm and the nature of scientific inquiry and methods undergo sudden transformations.

There are various philosophical schools that have roots in historicist tradition, among them are hermeneutics, pragmatism, social constructivism, semiotics, and activity theory/the cultural-historical school. Historicism became influential and dominant epistemology in the end of the 20th century.

**Pragmatism**

Pragmatism may be considered as one of the epistemologies belonging to the historicist family. According to pragmatism, it is important to examine the values and interests of each theory or knowledge claim and consider its practical consequences. The specific feature of pragmatism is the emphasis it puts on the purpose of research and cognition. Every theory and also every conception supports some goals and interests and should be also according to them justified. Knowledge is not
neutral and objective but always supporting some interests at the expense of other interests. Purpose and practical consequences are what matters, what should be the criteria for choosing to support some values and not others. Thus, what is valid knowledge is mostly evaluated from the perspective of what interests and goals this knowledge supports. Also the truth is viewed from the perspective of goals and consequences. It is always important to consider the practical consequences of regarding a particular theory as true.

**Importance of theories of knowledge**

Theories of knowledge, such as empiricism and positivism, have led to a number of assumptions about the world and reality; objects of research; users and their cognition; meaning, concepts and language; and other important issues in theories of information science. Every epistemology implies a view of how research should be done and what methods should be used in the research.

Whereas empiricism should not be confused with empirical research, it has impact on how empirical research should be done. A particular empirical study may be influenced by different epistemological positions that imply different methods and ideals of research. In empiricist and positivist tradition this ideal has been obtaining the facts, i.e. the observations on which different observers can agree on. The ideal has been to avoid the subjectivity of the individual researcher. Those results of research that can be reproduced are considered as truly scientific. The important underlying assumption is the view that observations and their descriptions are neutral; they are not influenced by the knowledge and perspective of the observer; and they are independent of cultural, social and contextual influences. Perception is seen as neutral reading of the objective reality that exists beyond the human mind. Researcher and reality are regarded as independent of each other.

Positivist and empiricist views are in contrast with more interpretative approaches, such as historicism or pragmatism, that view researcher and reality as inseparable; and knowledge of reality as constituted by the researcher's experience that is influenced by his knowledge and pre-understanding, and that takes place in a particular cultural and social context.

**3.4 Epistemologies in information science**

Various epistemological positions have been influential in various approaches and theories of IS, such as theories of information retrieval and understanding of information; theories of subject analysis and classification; or theories about users, their cognition and information seeking behavior [Hjørland, 1998].
The study of users and their cognition in IS was influenced by different psychological theories. Thus, as different epistemologies influenced theories in psychology, they also had impact on the studies of users in IS. Empiricism dominated psychology in the first half of the 20th century and is mostly associated with behaviorism. Rationalism was influential in the so called “cognitive revolution” that dominated psychology and was also very dominant approach in IS in the second half of the 20th century until the late 80's. Rationalist psychology saw users as born with some kind of psychological pre-understanding, or inborn concepts and categories which enabled them to interpret sensory data or observations. The “cognitive revolution” was closely related to cognitive sciences and artificial intelligence that compared human mind with a computer and tried to build computer systems according the way human brain works.

Historicism on the other hand regards knowledge as culturally and socially determined. It puts more emphasis on historical development of knowledge and concepts, study of contexts and evolutionary perspectives. The historical perspective on studying users and their cognition is concerned with how users and their cognitive structures are influenced by the culture and society in which they live; how concepts and knowledge developed in the history; how scientific theories influence the concepts and knowledge, while being influenced by language and other social and cultural and historical aspects. Historicism “implies a new view of users as social and cultural beings, and a more sociological-epistemological view on information seeking” [Hjørland, 1998, p. 610].

Various methods of classification and subject analysis may be also related to epistemological views. According to empiricism, the right method of classification is statistical analysis based on resemblance. In bibliographic classification, e.g., documents can be clustered based on the common terms in these documents. Rationalism, on the other hand, considers as a right basis for classification logical categories or divisions. This approach has traditionally been represented by facet analysis which is based on pure logical categories. In contrast to these two views, historicism promotes a classification based on the historical and social development of domains, scientific communities and their language and concepts.

According to pragmatism, classification should be based on the evaluation of its intended purpose and use. Different groups of users in various domains or theoretical perspectives may have different needs and interests. Classification is therefore not neutral and universal (as seen in rationalist, empiricist and positivist traditions) but it should always support interests of a particular user group and suit a certain purpose. Theoretical principles upon which a classification is built should reflect this. It should be clear...
how a particular kind of subject representation or classification meets certain needs and suits a given purpose.

Various epistemological and metatheoretical positions to LIS, that have been addressed in the special issue of the Journal of Documentation [2005, vol. 61, no. 1], include: critical realism; pragmatism, neopragmatism and sociocultural theory; phenomenology; structuralism and post-structuralism; constructivism, collectivism and constructionism; hermeneutics and classical positions empiricism, rationalism and positivism.

One example from the current research in LIS related to positivism is the studies of inter-indexer consistency. Their methodology has been positivist in its orientation on counting the agreements of the indexers rather than uncovering the reasons for why they index the way they do, and rather than focusing on the documents themselves. More importantly, the assumption that indexing is a value-free objective process is very strongly related to positivism. The pragmatic criticism of this assumption regards indexing as “an act of making priorities in mediating content for potential users” [Hjørland, 2005a, p. 146] and thus an activity based on the values and purpose that can not be neutral.

3.5 Knowledge organization and theories of knowledge

In the context of this thesis, it is most of all important to consider what role theories of knowledge play in various approaches to knowledge organization, i.e. how they influence the views of how classification, indexing and subject representation should be done and how KOS should be developed.

Broughton et al. [2005] identified the following theoretical approaches to KO: 1. traditional approaches to KO; 2. facet-analytical approach; 3. information retrieval as a challenger and closely related discipline to KO; 4. user oriented and cognitive views; 5. bibliometric approaches; 6. domain-analytic approach; 7. other approaches. Hjørland [2008b] discusses these approaches in detail and demonstrates how they may be related to theories of knowledge. His analysis suggests that empiricism, rationalism and positivism are theories of knowledge that have been very influential in information science and that are still alive in much of theoretical and practical work within the field. Other theories of knowledge related to historicism and pragmatism are according to the author more fruitful and could contribute to the advancement of the field.

1. Traditional approaches

Traditional approaches include traditional large classification systems used in libraries, such as DDC, LCC and UDC. Each of them reflects different principles behind their construction and especially DDC
was mostly developed for library management purposes and is thus considered rather non-academic, and not suitable for developing theoretical principles of KO. Traditional approaches reflected the general belief in librarianship that classification should be mostly dependent on practical issues. KO has been seen as a syntactic, rather than a semantic activity. No need was recognized for incorporating analysis of meaning and interpretation in these practice-driven activities, not to speak about formulating the theory of the field. Among the principles characteristic for the traditional approaches of KO are the principle of controlled vocabulary, rule about the most specific description attributed to Cutter, principle of literary warrant coined by Hulme and principle of organizing from the most general to the most specific.

2. facet-analytical approach

Facet-analytical classification founded by S. R. Ranganathan used five broad logical categories, called facets, according to which all subjects were analyzed and classified. It relied on the use of the analytico-synthetic method, according to which a subject of a document was first divided into basic concepts (analysis), and these concepts were than combined to describe the subject (synthesis). Ranganathan's approach represented a competing view to enumerative, hierarchical classification systems, such as UDC or DDC. It was not enumerative and pre-coordinate. This allowed for better flexibility in updating the system and it also enabled an item to be classified from more perspectives, by assigning more facets to it. Ranganathan's classification was developed into more advanced system by the British Classification Research Group.

The underlying assumption in the facet-analytical approach was that all subjects can be analyzed according to the facets which were logical categories, neither developed empirically, nor related to historical and social contexts and literatures of different domains, is related to rationalist epistemology. The facets were universal categories developed to suit many purposes. This is related to another a positivist assumption is that meaning is context-independent and that KOS can be objective and universal.

3. Information retrieval

IR is a subdiscipline of IS competing with and also closely related to KO as they are both trying to solve the same kinds of problems: how to facilitate retrieval of documents relevant to the user query. IR has been criticized as a reductionist approach that has been oriented on technical solutions of the user query-document match, without considering theoretical foundations of the objects of its study.
While the distinction has been made between a “system-oriented” and “user-oriented” view in IR, the former focusing on developing the best IR systems with the most powerful algorithms, the latter one emphasizing users' information seeking behavior, Hjørland [2008b] points out that this distinction is unnecessary or even misleading. In the user-oriented tradition IR systems are built and evaluated based on the evaluation of ordinary users, whereas the system-oriented view has been based on the expert evaluation of recall and precision. Neither of the two approaches considers issues such as what should be investigated to improve various controlled vocabularies or how are the documents found in the information search related to different perspectives and theories within particular domain. Thus, IR has been strongly influenced by positivism as it “has mainly been based on statistical averages, and has neglected to investigate how different kinds of representation and algorithms may serve different views and interests” [Hjørland, 2008b, p. 92].

4. User studies

Approaches within KO that use empirical studies of users as basis for designing and evaluating KOS have been very popular in IR as opposed to system-oriented views mentioned above. Hjørland [2008b] argues that classification based on users' opinion must necessarily be a naive, amateur, pre-scientific classification. User-oriented views are based on the belief that users themselves are able to judge what is relevant (relevance feedback) and they can express their information need. User studies have been relying on amateurs' opinion while neglecting the study of documents.

More importantly, both user studies and approaches to information retrieval based on expert-evaluation are based on positivist assumption that a representation is objective and neutral and that it is possible to develop a universal representation that suits all purposes. IR is also positivist in that it views the subject of document as neutral property of the text that can be simply abstracted or derived from the text, usually using some lexical and algorithmic methods. In its user-oriented tradition, this approach is also subjectivist because it studied concepts as creations of an individual user mind. It nevertheless regards representations of the subjects of documents as objective and universal, one that could answer questions of many different users. It does not consider that it might be useful to represent the same document differently for different purposes, as would imply pragmatic epistemology.

5. Bibliometric approaches

Bibliometric approaches use bibliographic references to uncover the citing relations between authors within domain. They use methods of bibliographic coupling and co-citation analysis to produce
bibliometric maps that show the citing relations among different authors and journals within domain. They also produce citation indexes which are valuable tools that record what articles and authors have been cited by other articles and authors. The basic assumption in creating KOS based on bibliometric methods is that citation/reference in a scientific article actually implies subject relatedness of two articles.

Bibliometric approaches are based on empirical methods; they use bibliographic references from scientific articles to create atlases of science, to map the relations between authors in a domain. However, as pointed out by Hjørland [2008b] bibliometric analysis can not be claimed to objective and neutral because the journal selected for the analysis are can not be selected in an objective way. The selection will always be biased and made from a certain perspective at the expense of some authors or journals.

6. Domain analysis

An approach to KO related to pragmatism is domain analysis. According to this view, KO should be concerned with how the knowledge is produced and organized within the individual domains and scientific communities. Different communities of users have different interests and therefore may need a different KOS. Any KOS is also always made from a certain perspective and supports certain interests and goals and is thus more suitable for some purposes and less for others. KO can never be objective or neutral, it always supports certain values and purposes, and to meet the needs of a particular group of users. These views are are in direct contrast to the positivist understanding of KO as neutral and objective activity.

3.6 Paradigms in medicine

In every domain there may be identified different metatheories, paradigms and epistemological trends influencing the field. In medicine a relatively new theoretical trend is Evidence Based Medicine (EBM). EBM began as a new movement in medicine in the early 1990's. Since than, it has received a great deal of attention and critique. The main argument of EBM is that the medical practice and education should be based on the best available results of the current medical research. According to EBM movement the old traditional paradigm in medicine, where the understanding of pathophysiologic theory together with clinical experience was of primary importance, was wrong in many aspects, especially in that it didn't recognize the importance of basing the medical practice and education on systematic evidence of medical research [Cohen, 2004].
The critiques of EBM regard the movement as based on problematic assumptions and point out that medicine has always been based on the results of medical research and that this argument is nothing new and revolutionary. The main problematic areas of EBM are related to the following issues:

Most important argument against EBM is that it is a poor philosophic basis for medicine because it is based on empiricism. EBM promotes the evidence provided by experimental studies designed to minimize bias. The belief that scientific observations can be made independent of the biases of the observer is very strongly influenced by empiricism. EBM in fact regards the “evidence” more important for the clinical decisions than understanding the pathophysiologic theory. In line with empiricist tradition, “EBM ignores the essential interplay between observation and theory” [Cohen, 2004, p. 38]. The original conception of EBM was based on the application of the methods used in epidemiology to the practice of patient care. The emphasis EBM puts on quantitative and statistical methods may be one important source of the empiricist assumptions present in this approach.

Critics of EBM also point out to the fact that EBM defines “evidence” in very narrow way and excludes important sources of information. The randomized controlled trials and epidemiological and bio-statistical methods and ways of thinking are regarded as the best ways of gaining “the evidence”, whereas other methods important, such as observational and ethnographic studies, qualitative methods and information gained from clinical experience are seen as less useful and unimportant.

The critiques regard the EBM movement as a threat to the autonomy of the doctor-patient relationship and argue that for the treatment of the individual patient the use of statistical results from the medical research is not always useful and often these results are not even available.

Finally, strong argument of the EBM opponents is that while the movement's goal is to improve the quality of health care by basing it on the results of statistically valid clinical trials, there is no evidence that this is actually the case. Thus the decision is apply EBM principles is more a political decision and is not itself supported by convincing evidence [Cohen, 2004, p. 39].

This demonstrates that various metatheoretical assumptions are at play not only in information science but also in other fields such as medicine. Different theoretical approaches in medicine may be also related to the theories of knowledge discussed in the previous sections of this chapter. EBM, according to which medical decisions should be based on the empirical and statistical evidence while disregarding other important kinds of evidence and sources of knowledge, is based on assumptions that are related to empiricist and positivist ideals.
There might, however, also be large group of researchers and practitioners that believe that not every diagnose and every method of treatment may be supported by empirical and statistical evidence. This group may emphasize the importance of practitioners' and patients' judgment and it may have completely different understanding of concepts like illness or health care.

A paradigm in the biomedicine different from EBM may be more hermeneutical or historicist oriented in that it regards the development of a particular concept of disease within the society as important and according to which such a concept may be understood differently in different contexts and within different scientific theories.

A concept of illness may be understood differently according to different epistemological ideals. Empiricist view defines illness by observable and measurable conditions of a human body that may be changed by applying certain methods of treatment. What method of treatment should be used is decided according to the available empirical and statistical evidence available. The society may consider one condition more desirable than the other but in the medicine itself there is no understanding of illness as being bad or good, it is a value-free condition that can be treated or changed with corresponding methods supported by the empirical evidence. Rationalist view, on the other hand, provides a system view of a human body and in this view a concepts of disease is understood as a failure in the system. From pragmatic and historicist perspectives, a concepts of disease is understood in the social and historical context. This view considers as relevant information about how the concept developed in relation to different scientific theories and how it was influenced by cultural and historical context.

3.7 Implications for knowledge organization

The theories of knowledge are at play in every domain, including IS and biomedicine. In KO, epistemologies such as pragmatism and historicism have important implications for how KOS should be constructed. In their view, every KOS is theory-laden and purpose-oriented. Every thesauri or classification is always made from a certain perspective and may be influenced, often implicitly, by empiricist, rationalist, historicist or pragmatic ideals.

If we consider a development of KOS in biomedical domain, the implications of historicist and pragmatic view have different levels:

On the first level we have different paradigms within a domain of biomedicine that may understand and classify concepts differently. In the biomedical domain the ways of observing and defining a disease or testing a cure may differ in different paradigms and approaches. This understanding more be more or
less influenced by the ideals of empiricism, rationalism, historicism or pragmatism.

On the second level, there are biomedical vocabularies, thesauri and classifications made in biomedical informatics and IS. These KOS are often made without considering the theoretical trends, paradigms and epistemologies within a domain for which they are made. There is a common belief that information specialist can classify everything without direct contact with the field and without explicitly considering what is going on in the field. If on this level we had understanding of different paradigms and epistemologies within a domain, we would be able to produce more useful KOS that could serve better the researchers and practitioners in the field.

On the third level, every concept within a domain such as biomedicine is developed from some kind of discourse or point of view. Thesauri that could reflect the perspectives behind concepts would therefore be much more useful and in closer connection with the domain itself. The ideal KOS should be able to trace back the different perspectives within a domain and represent them on the concept level. However, such semantic tool is difficult to make as it requires the identification of the theoretical and epistemological trends in the domain. Both good subject knowledge and philosophical background is necessary to do this kind of investigation.

4 Concept theory and semantics in knowledge organization

4.1 Importance of semantic issues

Semantic issues are present in many research areas and problems studied by information science. In particular knowledge organization and information retrieval cannot avoid dealing with meaning, semantics and theories of concepts. KO is engaged in the design and evaluation of classifications, thesauri, semantic networks and other kinds of semantic tools that serve information retrieval and knowledge organization purposes. It is also concerned with the knowledge organizing processes such as indexing, classification, document description and subject analysis/representation [Broughton et al., 2005]. The core of these processes is that they describe, classify and represent the intellectual and semantic content of documents. IR systems rely on the use of various lexical matching techniques and algorithms, as well as various semantic tools that help to represent the intellectual content of documents and match it with user queries. In all these tasks and processes semantic issues play very important role. Concepts and meaning are behind the information needs and questions of users, they are related to the intellectual content of documents, and they are also present in the semantic tools that help to connect the two. The importance of concepts and meaning in information seeking process is also
apparent in the goal of the UMLS that aims at improving the conceptual connections between users and the information they need [Humphreys, 1993]. Given that IS and KO in particular study the information seeking process and develop KOS that represent the meaning and semantic relations of terms, it is quite surprising that the issues related to semantics and concept theory have been overlooked in our field.

4.2 Research on concepts

Concept theory and semantics are extremely broad fields of study, with many disciplines contributing and not much consensus. Semantics is the study of meaning and concepts are generally understood as the meaning of words therefore these two should be closely connected [Hjørland, 2007, p. 369] into one field. The study of concepts and meaning has been, however, scattered in many different disciplines and the study of semantics is often not integrated with the study of concepts. Concepts, meaning and semantics are studied in psychology, philosophy, linguistics, sociology, artificial intelligence, semiotics, discourse analysis, terminology, library and information science, cognitive science and other fields. This interdisciplinary character is one reason for the complexity and lack of consensus in the theory of concepts.

In the philosophy, different views of concepts often reflect different views of much broader topics, such as the study of mind, cognition and language, view of the world and philosophy in general. Concepts are regarded to be constituents of thought and meanings of words. They are essential for categorizing the world and for communicating about the world. There is, however, considerable disagreement about what concepts are, what roles they play in human reasoning and communicating and how they can be defined. Psychologists often regard concepts as internal representations that function as vehicles of thought in human mind or brain. In logics and formal semantics, concepts stand for sets of real and possible objects and their functions. Philosophy usually understands concepts as something distinct from the representations that express them [Rey, 1998].

In philosophy, the study of concepts goes back to classic philosophers such as Plato and Aristotle. In the last decades of the 20th century the psychological perspective on concepts has been dominating in the research on concepts. This view is related to the cognitive sciences, cognitive psychology and artificial intelligence. Cognitive tradition regards concepts as structures in the individual human mind and brain. It bases the study of concepts on the experimental studies of human cognitive processes. Cognitive view influenced the way concepts were understood in information science. Critiques of the cognitive viewpoint mainly to the fact that this view completely neglected the study of concepts in
scientific disciplines. Domain analysis was developed as an alternative approach to the cognitive view of concepts. It emphasized the importance of the study of concepts in relation to the theories and paradigms in a particular domain [Hjørland, 2009, p. 4].

Recent monograph that provides a broad overview of research on concepts in psychology is Murphy [2002]. The topics covered by the book such as “category learning“ and “conceptual developments in infants and children“ indicate the strong psychological/cognitive orientation. Murphy reviews a lot of empirical work on the study of concepts in psychology and relates it to three theories of concepts: (1) the prototype view, according to which there exist a prototype of each category/concept that we refer to when we encounter a new object; (2) the exemplar view, according to which the concepts are formed by remembering examples as we encounter objects, and (3) the theory view according to which concepts exist in theoretical framework with other concepts [Benzon, 2004]. Murphy is, however, inconclusive as for which of the theories is the correct theory of concepts. Regarding the empirical studies, Murphy concludes that much research has been done in developing models that explain a particular phenomena or issue but that do not provide a general theory of concepts valid or applicable to broader set of phenomena.

Theories of concepts

The major theories of concepts, regarding the identity and structure of concepts, as discussed in the literature [Margolis and Laurence, 1999 and 2005; Kavouras and Kokla, 2008] are:

- **The “classical theory” of concepts** (corresponding to the classical view in Murphy [2002]) according to which concepts have definitional structure, i.e. they are composed of simpler concepts that express necessary and sufficient conditions for memberships in a category/concept. Every object either belongs or does not belong in the category and all the members of the category are equal, no object is a better member of the category than other. Thus categories in the classical view have clear boundaries. Concepts are acquired by learning their definitional characteristics. Classical view of concepts has been heavily criticized both by psychologists, who claimed that many empirical findings (such as that some objects are viewed as more representative of a category than others) could not be explained by this theory and by philosophers who argued that there are very few concepts that can be identified by a precise definition and that concepts don't have the definitional structure.

- **Probabilistic theories of concepts** including the “prototype theory” that regard concepts as
having a prototype structure, i.e. the membership in a category is identified by the virtue of
family resemblance or similarity characteristics and some members are more typical of the
category than others. The prototype theory is mostly associated with philosopher Ludwig
Wittgenstein and psychologist Eleanor Rosch. Critics of the prototype theory argue that it is
problematic to apply on the more complex concepts and more reflective judgments where
similarity comparison is not sufficient explanation. One suggested solution to this problem is
the view that concepts are constituted partly by the prototype and partly by the conceptual core
that specifies the concepts in more substantial way.

- **Theory-based theories** of concepts that view concepts as related to one another in a theoretical
  framework in the same way as terms in a scientific theory. A concept is defined by its unique
  role in the particular theory. The human ability to categorize is not based on based on the
  observation of a set of perceptible features, but on them being aware of the *essence* or internal
  structure of the concept. Concepts are in the theory-based theories thought of as embedded in
  a person's knowledge about the world, i.e. their mental theories of the world. The problem with
  this view is that people may have an inadequate or incorrect mental theory about a concept and
  still posses a concept. It is also difficult to describe the essence of a concept.

- **Neoclassical theories** of concepts that adopted the classical theory of concepts to the modern
cognitive science and that define concepts only partially.

- **Conceptual atomism** that views concepts as having no semantic structure and that is mostly
  a negative theory about what concepts is not.

While the psychological perspective on the study of concepts has been influential, other important
alternative views of concepts have been overlooked. Hjørland [2009] puts forward what he calls “the
**post-Kuhnian view of concepts**” which indicates that the connection between scientific paradigms and
concepts is recognized in this view. Whereas the dominating psychologistic and individualist paradigm
in the study of concepts considers concepts as developed in the individual mind, in the post-Kuhnian
view the development of concepts is regarded as mostly influenced by different theories and paradigms
in science, and generally developed in the social, cultural and historical context. This view is also
connected to perspectives such as pragmatism, activity theory and hermeneutics which have been
neglected by the main-stream researchers in the study of concepts. Post-Kuhnian view is a historical
and theory-oriented approach to the study of concepts. It seeks the relation between different theories
of concepts and major epistemological positions. The important claim of the post-Kuhnian understanding of concepts is that the way concepts are defined by scholars in different fields is related their methodological and epistemological ideals. This view of concepts has important implications for how processes such as subject analysis, document description and indexing should be done and how KOS should be constructed.

**4.3 Concepts and semantic issues in knowledge organization**

In KO, semantic issues are mostly related to defining and representing concepts and semantic relationships in KOS. The relation between KO and concept theory/semantics may be understood by considering the KOS as semantic tools representing a certain selection of concepts and information about their semantic relations [Hjørland 2007 and 2009]. KOS such as thesauri or controlled vocabulary are organized collections of terms, in which some terms are represented as preferred terms or descriptors and some as non-descriptors. Terms are also usually put in some kind of semantic structure, usually hierarchical, from which semantic relations can be inferred. Concept is commonly defined as a meaning that can be identified or expressed by the class of synonymous terms.

Semantic relationships are the relationships between concepts, i.e. between meanings. This is not the same as the relationships between terms or words as one concept may be represented by various terms but also by a numerical notation. Also, terms may have few different meanings; they may refer to more than one concept. Thus, concepts not terms are at the core of semantic relationships. All KOS address in some way how concepts represented in them are related to each other. Main kinds of subject relationships are equivalence, associative and hierarchical relationships [Green, 2001].

Concept is usually recognized as important and essential unit in KO and IS in general. The role of the concepts in KOS was acknowledged in the recent changes introduced in the MeSH, probably influenced by the UMLS. Whereas the historical structure of MeSH emphasized the relationships on the term level, the thesaurus was redesigned to become a more concept-oriented. Before the change a descriptor in MeSH was viewed as a group of terms, where the preferred term was the name of the Main Heading. In the current view, a descriptor is understood as consisting of one or more concepts. In the descriptor class one concept is regarded as most important, as a preferred concept. Other concepts are subordinate or “narrower” in meaning [Nelson, 2001].

Various theories of concepts have different implications for how KOS should be constructed, evaluated and used, and the classification and subject representation performed. Principles and methodologies for
building KOS should therefore be based on the study of meaning and concepts. This has, however, often not been the case. Semantic tools are often constructed without explicitly considering what view of semantics and concepts they rely upon. The importance of semantics and concept theory has been neglected in IS. Most books about organization of knowledge, indexing subject analysis and representation do not consider the theories of semantics and concepts [e.g. Cleveland and Cleveland, 2001; Taylor and Jourdney, 2009; Rowley and Hartley, 2008]. Some of them do not even mention concepts at all. The reason for this may be the fact that theories of semantics and concepts are often not clear about how should concepts and semantic relationships be defined and represented in KOS and what should be the methodological and theoretical basis for the construction of KOS.

Various approaches to KO deal with semantic issues in different ways, often influenced by epistemological assumptions discussed earlier [Hjørland, 2007]. Traditional classification systems, such as UDC or DDC, are based on the principle of literary warrant and thus find the semantic relations in the scientific literature. This has however been based on the positivist assumptions that the literature contains facts reflecting objective knowledge and that also subject representation of the literature can be objective and universal. Facet-analytical approach influenced by rationalism based KOS on a priori semantic relations derived by logical principles. In IR semantic relations have been viewed as statistical relations between words and documents. It has been assumed that subject is the property of the text of a document that can be determined by automatic methods. User and cognitive studies based the semantic relations on the empirical studies of users and neglected the study of literatures and domains. Specific semantic relations have been introduced by bibliometric approaches. These approaches regard two documents as semantically related when they cite each other, when they are cited together by other document or when they are bibliographically coupled. According to Hjørland [2007] these meaning relations are social because they reflect social behavior in scientific communities whereas those discovered by different scientific theories and also traditionally represented in KOS are ontological relations. Domain-analytic approach considers semantic relations as determined by different theories, paradigms and epistemologies. It bases semantic relations in the scientific literature such as in the case of the principle of literary warrant. It does not, however, regard semantic relations as objective and universal but as influenced by various theories and views that are present in every domain.

Main theories of knowledge imply their own view of concepts and semantics. This view is discussed in the following section.
4.4 Theories of knowledge and their view of concepts

Theories of knowledge that have been described in chapter 2 imply beside a particular methodological ideal also a certain view of concepts and semantics [Hjørland, 2009]. In different theories and approaches to concepts it is possible to identify various epistemological assumptions that correspond to empiricism, rationalism, positivism, historicism, etc. Also the way concepts are defined by scientists is related to their epistemological and methodological ideals. Hjørland [2009] is an overview of the views of concepts related to the main epistemologies. His work suggests that an important unifying view of the various theories and approaches to semantics and concepts, scattered in the literature of many disciplines, may be considering their relation to the theories of knowledge.

Empiricist view of concepts

According to empiricism knowledge is based on observations that are “given”, not contextual or theory-dependent. This has also influence on the way concepts are defined. Empiricism considers concepts as based in perception, i.e. based on the resemblance among things and on the observable features. Concepts and their representations are learned, not innate as in rationalist view. When we learn a concept, we learn about the similarities between things and about conventions between things and words. Empiricist ideal is to avoid theoretical selections and interpretations of observations and this also applies to concepts. Empiricist view defines a concept based on observable features but fails to argue what characteristics or properties are the most important for defining a concept and why. The criticism of this view, however, argues that it is not possible to define concepts without any theoretical understanding as well as it is not possible to create atheoretical classification.
In biology, a classification based on cluster analysis, also known as phenetics, classifies species based overall similarity, i.e. on observable features shared by members of taxonomic unit [Ereshefsky, 2001]. Phenetics as an attempt to develop theory-neutral universal classification subscribes to empiricist epistemology. Classification should only be based on observing the resemblances in the nature and should avoid theoretical selection of features observed [Ereshefsky, 2001, p. 61]. Cluster analysis as a method based on grouping the objects into clusters based on similarity measures was criticized by Cooper [2005]. According to her it is not possible to construct atheoretical classification using this method because a scientific theory is needed to select the variables that measure the true properties of the objects. If the variables are poorly chosen, the cluster solutions will be worthless. And to decide what variable/property is relevant requires a scientific theory. Thus, the empiricist ideal of avoiding theoretical selections and interpretations is problematic.

**Rationalist view of concepts**

Rationalist epistemology emphasizes a priori knowledge, theorizing and logical models to observation and experience. Classification based on “logical division” is a method influenced by rationalism. In rationalist view, concepts are a priori logical structures that exist in the human mind before the sensory experience. Concepts are viewed as “logically exclusive and mutually exhaustive classes constructed on logical principles rather than based on observing reality” Hjørland [2009, p. 20]. Rationalist view of concepts was influential in the cognitive revolution. In psychology, much research focusing on innate domains of knowledge and concepts continued in the rationalist tradition.

In biological systematics, method of classification based on logical division, according to which a concept of species should be defined on the basis of the “essential characteristics” shared by the members of the species [Ereshefsky, 2001, p. 95], is related rationalism [Hjørland, 2008a]. Another example of rationalist view of concepts is a method of data analysis called “formal concept analysis” (FCA) that has been applied in many disciplines including software engineering, artificial intelligence, information retrieval, linguistics or psychology [Priss, 2006]. Concepts in FCA are clusters of formal objects paired with clusters of formal properties. Formal objects and formal attributes may be e.g. documents and terms, or words and meanings. FCA is thus mathematical and formal method of identifying formal concepts corresponding to rationalist epistemology which defines concepts as a priori logical structures. Hjørland [2009] argues that natural kind concepts cannot be considered as formal objects and thus other views of concepts are needed. As an example he mentions the concept of species in biology. Definition of such a concept must be based both on empirical research and various
theories in the field of biology and can not be established formally.

**Concepts in logical positivism**

The philosophy of logical positivism combined both rationalist and empiricist assumptions. Logical positivism reduced concepts to combinations of sensory attributes. A theory of meaning related to logical positivism is operationalism [Svenonius, 2004]. In this theory, scientific concepts are defined operationally, i.e. they are defined through operations by which they are measured. As an example of operationalist definition Svenonius mentions the precision-recall measure developed to measure the degree to which an information retrieval system is able to retrieve only documents relevant to the query and degree to which it retrieves all the relevant documents. Another example is the subject of a document that was defined according to the methods of measuring the frequency the words in the document.

Thus, according to empiricism and rationalism (and positivism) concepts come either from senses and experience or they are inborn structures in the human mind. Concepts and cognitive structures are biologically determined. All these views disregard social and cultural influences on the development of concepts, they ignore that human cognition is embedded in the sociocultural context. This has been called “the empiricist-rationalist trap”.

In contrast, theories of concepts related to historicism, hermeneutics and pragmatism represent a social turn in understanding the human cognition [Hjørland, 2009]. According to these theories of knowledge, human cognition is culturally and socially determined, i.e. influenced by our language and symbolic systems, by our pre-understanding and conceptualizations.

**Historicist view of concepts**

According to historicist epistemology, observations are theory-laden and culturally and socially influenced. Knowledge about social contexts and historical developments is considered as relevant. Similarly, concepts are seen as meanings that developed historically and that are related to discourses, perspectives and paradigms in domains. To learn a concept means to learn about how the concept is understood in our culture and how this understanding developed in the history. Thus, conceptions are relative to a particular culture and discourse. The view that concepts develop in the social and cultural context and they are influenced by our theories and pre-understanding is quite different from the empiricist and rationalist understanding of concepts as biologically given, either derived from experience or inborn.
Pragmatic view of semantics and concepts

Pragmatism, similarly as historicism, regards concepts and meanings as developed in social and cultural context. Concepts are in pragmatic view defined functionally, with respect to the purpose. Concepts are viewed as representing functionally equivalent classes of things. What is to be equivalent is decided according to the particular purpose and theoretical perspective [Hjørland, 2009]. Concepts are seen as related to specific tasks and serving certain goals. This also implies that different purposes may require different concepts.

According to pragmatist view all descriptions and conceptions of objects are theory-laden, i.e. they are always made from a certain perspective and serve certain interests. Concepts and semantic relations are not neutral, but theory-laden and context-relative. They depend on our understanding of the world and on our theories. They are relative to the social, historical, cultural and domain-specific context. They are shaped by scientific theories and paradigms. If there is some paradigm in the domain that is dominant, it is likely that also concepts will be influenced by this paradigm. However, the alternative paradigms and approaches may develop alternative view of concepts driven by different interests and different theoretical position.

Implications of the post-Kuhnian view

The implications of the post-Kuhnian view of concepts and semantics is very important and fundamental. Whereas traditionally it is has been assumed that concepts and semantic relations are objective and neutral and the role of KOS is to represent and organize them in objective and universal way, this new perspective sheds on these issue quite different light. Concepts and meanings are linked to different views and interests and thus also KOS must necessarily support some views and interests at the expense of others. Ideally, the KO should uncover the different paradigms in the literature of a domain and should develop KOS that represent concepts and semantic relations in relation to these paradigms and views. All attempts to create an objective KOS and objective view of concepts and meaning, and attempts to standardize terminology, must necessarily result in suppressing some views and in losing the context in which concepts and meaning are situated. KO must therefore consider various ways of defining concepts and determining semantic relations and decide what view of concepts and semantics is most fruitful for a given purpose.
This view is often not considered when constructing KOS. The principles of KOS construction are mostly based on positivist assumptions that scientific literature contains objective facts and that the meaning and concepts used in the literature may be captured in KOS in an objective and neutral way.

5 Understanding the meaning in the UMLS

5.1 Biomedical vocabulary integration

The UMLS is a large project of biomedical vocabulary integration in which biomedical concepts play a key role. Its main components, metathesaurus and semantic network, are semantic tools developed to aid the retrieval and integration of biomedical information. They were created in response to the diversity of terminology that exists in the biomedical domain and that makes the efficient retrieval and integration of biomedical information difficult and problematic task. The metathesaurus is a selection of biomedical terms organized into concepts, and it provides a rich network of semantic and lexical links among them. The semantic network provides a unifying semantic structure to the metathesaurus. It defines broad meaning categories called semantic types into which concepts are assigned, as well as a variety of semantic relationships plausible between them.

Biomedicine is a very diverse field, with many different specialties and disciplines. These specialties may develop their own language, and rely on their own methods and literature. However, in the same way as every part of human body is not functioning in isolation but as a part of the whole organism, researchers and healthcare practitioners in one specialty (e.g. chirurgy) need to know about the contributions of other specialties to their area of research. Also much of the research has become interdisciplinary.

In the biomedical domain different KOS have been developed for different purposes. Some of them may be large classifications of diseases developed for statistical purposes (such as ICD - International Statistical Classification of Diseases and Related Health Problems), other are much more specialized or serve other practical and managerial purposes within health care. All these classifications and vocabularies are semantic tools, they provide a certain selection of terms referring to some concepts. The terms and concepts represented in each vocabulary may differ. Each terminology or vocabulary is based on some theoretical assumptions and beliefs concerning the concepts used in them. These assumptions may not be explicitly stated but they nevertheless influence the way concepts are represented in the vocabulary, as well as what terms are chosen as preferred terms, what relationships are considered hierarchical, or what terms are considered synonymous.
Health care practitioners and researchers need to have access to vast range of information sources. The UMLS that integrates the various biomedical vocabularies into one common KOS must be seen as an attempt to provide this integrated access to diverse biomedical sources.

5.2 Epistemological perspectives on the UMLS

The way meaning, concepts and semantic relationships are understood and represented in the UMLS metathesaurus is the primary research interest of this paper. While various problems in the integration of vocabularies into the metathesaurus caused by the differences in granularity, scope and intended use of the individual source vocabularies have been discussed in the literature, e.g., [Bodenreider, 2001], the theoretical problems inherent in the integration of vocabularies into the metathesaurus received no serious attention. These issues are related to broad theoretical principles of how KOS should be constructed and how meaning and concepts should be represented in them.

Thus we may ask what principles have been applied in constructing the UMLS? Although there is a huge amount of literature about the development of UMLS, its application and use, most of the research has been oriented on specific technical problems of vocabulary integration and UMLS development and no serious attention has been paid to the general principles and philosophy behind the project. This corresponds to the situation in KO in general where KOS are often developed without explicitly considering the principles behind their construction and not much attention is focused on developing a general theory of the field.

Multiple meaning representation

The UMLS metathesaurus represents biomedical concepts derived from various terminologies. However, the meaning is the UMLS is often quite different than in its source vocabularies. This includes to view of concepts, synonymy and hierarchical relationships. While the semantic information provided by the source vocabularies is an important source for the UMLS, the metathesaurus own view of meaning is often quite different.

Concepts are the most important components of the metathesaurus structure, through them are the various terms linked and integrated. The role and function of concepts in the metathesaurus is the main controversy, as well as main achievement, of the UMLS. Concepts in the UMLS are clusters of synonymous terms taken from source vocabularies. Synonymy is most important semantic relationships in concept-based terminologies, including UMLS. Synonyms are terms equivalent in their meaning.
However, it is common to find terms with closely related meaning that can be considered as
synonymous under certain circumstances, i.e. in a particular context. These “relative synonyms” may
be treated as if they named exactly the same concept in some controlled vocabularies. This is
determined by the granularity of the particular vocabulary, by its scope and use.

When a source vocabulary is integrated into the UMLS, the concepts and concept names available in
the source vocabulary must be organized in relation to the existing UMLS concept structure. This is the
most important part of the editing work. The concepts and concept names are first algorithmically
matched with the UMLS content. Next, the result is reviewed by human editors. Because the source
vocabulary may have different view of synonymy than the UMLS, the two views must be aligned
[Fung et al., 2005].

This was also the case in the integration of Systematized Nomenclature of Medicine - Clinical Terms
(SNOMED CT), the largest vocabulary that has been integrated into the UMLS. In the UMLS,
SNOMED concepts were merged for various reasons. Sometimes concepts that were considered as
non-synonyms in SNOMED CT because of the very fine granularity of the vocabulary, such as concept
names of a medical product differing only in their packages and sizes, were in the UMLS merged into
a single concept. Another reason for merging SNOMED CT concepts involved concepts containing the
“Not Other Specified” qualifier. Also, concepts that were non-synonyms in SNOMED CT because of
its strict separations of hierarchies, such as the concepts “stab wound (disorder)” and “stab wound
(morphologic abnormality)” that in SNOMED CT belong to different hierarchies (“body structure” and
“clinical finding”) were in the UMLS merged together into one concept. The reason for this was that
although these two are “theoretically different concepts, in most clinical situations this distinction is
neither necessary nor helpful” [Fung et al., 2005, p. 488]. Some cases of missed synonymy in
SNOMED CT were also discovered during the integration.

In another situation, concepts that were considered as synonyms in SNOMED CT were split into
multiple concepts in the UMLS. This decision was made by the human editors in cases where they
found the “difference in meaning substantial enough for a particular concept name to be split out and
put into another UMLS concept” [Fung et al., 2005, p. 488]. For example “motor vehicle accident“ and
“motor vehicle accident victim” which were considered as synonyms in SNOMED CT were split and
the former was put into another existing concept in the UMLS “traffic accidents” while the latter
formed a new concept.
The purpose of the UMLS is to link various source vocabularies. Thus, the UMLS concept view is a unifying view through which contents of various vocabularies are linked together. “However, this does not imply that the UMLS view is the only 'correct view' of synonymy” [Fung et al., 2005, p. 489].

The UMLS is trying to preserve the original content of its source vocabularies and the particular source's view of synonymy can be retrieved from the metathesaurus. In the UMLS own view, synonyms (all atoms with the same CUI) are grouped together to form a single metathesaurus concept. The source view of synonymy is also preserved and represented by an atom level relationship.

What is then the relation between the meaning in the UMLS and in its source vocabularies? This issue has been addressed by Campbell et al. [1998b] who claim that the distinction between two types of meaning, i.e. extension and intension, as described by the Ogden and Richards semiotic triangle, can be used for understanding the meaning represented by concepts in the UMLS metathesaurus in relation to the meaning of terms of its source vocabularies. In the semiotic triangle, the extensional component of an expression is the set of physical objects to which expression refers, whereas intensional meaning is the set of characteristic features of the object used to identify it. According to Cambell et al. [1998b] the metathesaurus provides a framework where the intension of terms can be preserved and their extensional meaning is unified. They argue that the meaning of a concept in the UMLS (represented by CUI) must be understood extensionally, by examining the characteristics shared by all abstract concepts (intensional representations) linked by the CUI.

The intensional meaning of a term in the source vocabulary is represented by its position in the structure of the vocabulary, by what terms are regarded as synonyms of the term and by other semantic relationships in which a term appears. This intensional meaning is context-dependent, i.e. it is different in various source vocabularies. This is given by the different purposes for which vocabularies were developed. This doesn't mean that the intensional meanings of phrases in one vocabulary, e.g. SNOMED, are more correct that the intensional meanings of another vocabulary, say MeSH. The goal of the UMLS is to represent the different intensions of the source while providing integration of the extensional meaning.

This leads us to an important feature of the metathesaurus that has been discussed above, i.e. the fact that metathesaurus to a great extend preserves the original meaning of its source vocabularies. It represents the semantic relationships (parent/child, siblings, synonymy) as they are understood in the source vocabularies together with the metathesaurus own view. Thus, it in fact represents multiple views of meaning derived from the source vocabularies plus a unifying metathesaurus view. The users
of the UMLS Knowledge Source Server may therefore see how the particular term is represented in a particular vocabulary, e.g. Medical Subject Headings. They also see the overall view provided by the metathesaurus, i.e. the concept entry to which the term has been assigned. Therefore what is provided in the metathesaurus is a view of multiple KOS.

Epistemological perspective
As discussed above, different views of concepts and semantic issues in KO are related to epistemological and methodological ideals that underlie theories in KO. According to the “post-Kuhnian” view of concepts which is related to historicist and pragmatists epistemology, concepts and meaning are relative to the context [Hjørland, 2009]. More specifically, concepts develop in relation to various theories, paradigms and approaches that are at play in every domain and that may develop their own competing conceptions and meanings to support different interests. Thus, the differences between understanding or definition of a particular concept in two subdisciplines or two paradigms in a domain may be functional because concepts are related to the various theories and serve certain interests. A concept is from a pragmatic point of view understood as a class of things that are functionally equivalent. What may be regarded as functional in one theory or according to one perspective may not be functional from another perspective.

The pragmatic-historicist understanding of concepts implies that different paradigms and perspectives in the particular domain may need different knowledge organizing systems, developed to support their specific needs. Pragmatic approach regards KOS as semantic tools created for a particular purpose and every selection of concepts and representation of meaning as reflecting certain perspective and theoretical position, and supporting some interests at the expense of other interests. Thus, a valuable KOS should explicitly reflect on what theoretical principles and views it has been built, for what purpose and with what understanding of concepts and meaning. In ideal situation, KOS should be able to uncover the paradigms and views in a domain and inform a user about them in the knowledge representation of the literature of the domain.

Pragmatic critique of UMLS
If concepts (meanings of terms) and semantic relationships are context-relative, the project that integrates various KOS into one system is highly problematic because it disregards that the differences in KOS may be functional, and, more importantly, that in the biomedical domain there may be different paradigms and perspectives that develop their own view of concepts and that these perspectives should
be uncovered in KOS.

From the historicist and pragmatic perspectives, in the theoretical and philosophical basis of the UMLS there is a lack of understanding of the history of concepts and of the theories behind the concepts in the individual KOS. The metathesaurus does not reflect the conceptual change within the subdisciplines or perhaps across the subdisciplines. In this view, the metathesaurus would be more valuable if it reflected the dynamics of change in meaning and the historical development of concepts.

**UMLS is “pragmatic”**

From the pragmatic perspective KOS should be constructed with a particular purpose in mind, as well as concepts defined in the way that matches particular interests. One could claim that UMLS meets these requirements. Its purpose is to support information retrieval and integration of various information sources in the biomedical domain. For this purpose, it seems useful to map different terminologies into a common framework and connect terms with possibly same or related meaning together. UMLS is a powerful tool for information retrieval in the domain as broad as biomedicine. Moreover, in its unique way of representing multiple views of meaning, UMLS is in fact uncovering the differences instead of blindly unifying the terminologies. The efforts the preserve the original meaning and to represent to differences between biomedical vocabularies must be acknowledged.

For the purpose of mapping the biomedical vocabularies into a common framework, UMLS defines a concept in a very unique way, as a cluster of terms used in different KOS. The meaning of a concept in the UMLS consists of all the meanings of terms from various vocabularies put together.

**Is UMLS based on the empiricist/positivist view of concepts?**

Is it than possible to regard UMLS, which from the historicist-pragmatic perspective ignores the view that concepts are context-relative, developed in relation to scientific paradigms and theories, as based on the empiricist or positivist view of concepts and knowledge organization?

The empiricist view considers concepts as based on the resemblance among things (such as in the case of “cluster analysis” mentioned before) and tries to avoid theoretical selections and interpretations. Let's consider if this view of concepts applies to the UMLS. It seems that it does not. Despite the fact that the concepts in the UMLS metathesaurus are “clusters” of synonymous terms taken from various vocabularies, and despite that the synonymy are often partially discovered by using lexical matching techniques which are based on the resemblance among terms, it is not possible to compare this approach with the method of cluster analysis, which identifies concepts on the basis of similarity
measures and which is related to empiricism. Clusters that form concepts in the metathesaurus are based on a particular view of synonymy and, although this view is not explicitly expressed in the literature, they are not claimed to be atheoretical. Terms are assigned into the cluster because they are either considered as synonyms in the particular vocabulary, or because they are regarded as equivalent in their meaning by the editors of the UMLS. The lexical matching techniques are only much cheaper and faster way of discovering possibly related and synonymous terms. The view of meaning is, however, based on the editors' judgment. The fact that the editors are domain experts reflects the view that the knowledge/theory of the biomedical domain is required to provide the meaning judgment in the UMLS. Another factor that influences the UMLS view of meaning is its purpose. For example synonymy are defined loosely on purpose, because it suits the goal of the UMLS to integrate terminology from various biomedical vocabularies.

The UMLS view of meaning

What then is the UMLS view of meaning and how it is determined? There two most important factors at play in this issue. First, the view of meaning is influenced by what is useful for the purpose of the UMLS, i.e. for the integration of biomedical terminology into a common framework. Second, the semantic information in the UMLS is based on what is right from the biomedical point of view, i.e. what is correct according to current scientific theories.

There seems to be a confusion in the literature about if there is a single correct view of meaning that should be represented in the UMLS. On one hand, the UMLS recognizes that different views of synonymy and other semantic relationships in the source vocabularies are related to the different purposes for that the vocabularies were developed and that they may be functional and useful in some situations. It does, however, maintain the view that some view of synonymy is the true view. Synonyms from the source vocabularies are evaluated and if they are not true synonyms, they are not included in the concept entry [McCray, 1995, p. 194]. In some cases, synonyms from the source vocabulary are split into two different concepts (e.g. “septic arthritis” and “infectious arthritis”, which are synonyms in MeSH, are in the UMLS established as separate concepts).

This indicates the assumption that there is some view of concepts and semantic relationships that reflects the true biomedical knowledge and reality and that it is the aim of the UMLS to represent this view. What meaning is the “correct” meaning is judged by the UMLS editors who are domain specialist. The view that concepts and meaning in the UMLS are based on the current scientific theories
and definitions of concepts such as diseases is demonstrated on the following example: The terms “hepatoma” and “hepatocellular carcinoma”, which are regarded as synonyms in SNOMED, are hierarchically related in the UMLS because “hepatoma” is defined as any tumor in the liver, and the “hepatocellular carcinoma” is its most common histological type.

We may therefore argue that the view of meaning in the UMLS is based on the positivist assumption that scientific knowledge is an objective and neutral representation of reality and natural order. The UMLS view of meaning, concepts and semantic relationships seem to be based on the current scientific biomedical knowledge in the sense that the semantic information represented in the metathesaurus and semantic network is supposed to reflect what is right in biomedical sense.

On the other hand, the UMLS recognizes that synonymy and other semantic relationships are context-relative and it represents the views provided by the source vocabularies, as they may be functional in certain contexts.

Fung et al. [2005, p. 489] explicitly state that while in the integration of SNOMED CT into the UMLS the two views of synonymy must have been aligned, this does not mean that the UMLS view is the only correct view of synonymy. And Campbell et al. [1998b, p. 428] argue that the UMLS recognizes the need to represent the notion of meaning, in their interpretation intensional meaning, that is “relative to the scope, granularity, context (hierarchy), synonyms, and annotations of the source terminologies”.

One could ask if the semantic types as broad meaning categories assigned to all concepts in the metathesaurus are not something like logical categories or facets and thus related to rationalist perspective. This is, however, not the case. The semantic types and relationships in the semantic network were developed empirically, based on the analysis of the source vocabularies included in the UMLS and on experiments using the UMLS test collection of queries and MEDLINE citation records [Lindberg, 1993, p. 285]. It is also recognized that the semantic types and many of the semantic relationships in the network are domain specific and they are viewed as representing the biomedical “common sense”.

Another factor influencing the UMLS view of meaning is the purpose of the UMLS itself. The UMLS links the various vocabularies in biomedicine, enables translations and mapping among them and generally serves the retrieval and integration of biomedical information. The concepts and meaning in the UMLS are also defined according to this purpose. The view of concepts as classes of synonyms taken from the source vocabularies is tailored just for this purpose. The definition of synonymy is loose
so that all the terms found to be useful to belong to the concept could be assigned there.

From the historicist/pragmatic perspective, another objection one could have against UMLS is the fact that it is based on existing KOS and not directly on the literature of the domain. The source vocabularies are already interpretations of the biomedical knowledge, often develop with different assumptions and from different perspectives. The reasons for that UMLS is built upon existing KOS are first of all economical. It is simply much faster and cheaper to use terminology that has already been captured in KOS and link it together. It is recognized that the UMLS should be based on the current use of the language and that it should reflect the domain knowledge. It is, however, not recognized that the study of different paradigms and perspectives in the literature of the domain would reveal another important kind of knowledge, the knowledge that is essential from the pragmatic and historicist perspectives.

We may conclude that the meaning in the UMLS is based on particular view that is believed to reflect the biomedical reality. While the differences among the source vocabularies and their view of meaning are recognized and represented, there is, however, a belief that the UMLS should be such a neutral and objective representation, based on the scientific theory.

The purpose of the UMLS is to enable the health care researchers and practitioners to find biomedical information from a variety of sources. The UMLS is therefore oriented towards its practical use. Researchers and practitioners need to have access to information that is valid according to current scientific theories. They need to find articles about how a particular disease should be cured according to the present scientific theories which are considered to be true and objective, or, at least, the best available. For the purpose of philosophical research, it would be interesting and useful to have a KOS that would represent all possible meanings of concepts related to various perspective and paradigms within a domain. A historical perspective uncovering the development of understanding diseases would be also interesting. From the perspective of the scientist and practitioner in biomedical domain it is nevertheless more important to have access to the best current available knowledge. Natural sciences have always been strongly influenced by positivist thinking. Even today most researchers believe there is objective truth out there that sciences must discover. And while understanding of some theory and phenomena certainly changes during the course of history, the latest theories are usually viewed as more advanced and closer to discovering the fundamental truth than their predecessors.

Positivist assumptions are therefore present in biomedical science and this is also obvious in the way
biomedical knowledge is represented in the UMLS. The view that there are objective scientific theories provides sufficient justification for creating a system like UMLS that helps to uncover meanings that are right and useful according to best scientific knowledge.

Pragmatist views are also present in the UMLS, although implicitly. Concepts are in the UMLS defined and meaning is represented with regard to the purpose of the system and to the practical consequences. In practice, it is useful to join various perspectives of meaning and to integrate biomedical terminology in the metathesaurus. For this purpose, the UMLS view of concepts seems to be useful. The UMLS, on the other hand, also represents the differences between various KOS and their views of meaning because they are functional and can not be ignored if the UMLS is to be of practical use.

6 Conclusion

Theories of knowledge play important role in various theories and approaches in information science. Epistemologies such as empiricism and positivism underlie the theoretical assumptions maintained by approaches to knowledge organization. Different theories of knowledge imply different methodological ideals and have profound influence on the views of how the work in our field should be done, e.g. how classification and subject representation should be performed. Not only various epistemologies imply a certain methodology in the field, they have also enormous impact on how the research process itself is viewed, what is regarded as relevant for the field to study and what is perceived as its goal. A good example is the difference between how the process of knowledge representation is viewed in epistemologies related to positivism and empiricism on the one hand, and pragmatism and historicism on the other. While the former argues for the quest of objective, universal and neutral knowledge representation and knowledge organizing systems that reflect the true reality and natural order, the latter regards knowledge organization as necessarily biased and perspective/theory-laden, supporting some values and interests, and influenced by sociocultural and historical context.

In this thesis we tried to use the epistemological lenses to look at the unique knowledge organizing system developed in the field of biomedicine, namely the Unified Medical Language System. We tried to question the theoretical and philosophical basis of this remarkable system. Mostly, our attention was directed towards the way how meaning, concepts and semantic relationships are understood and represented in the UMLS metathesaurus. To uncover the theoretical assumptions behind the construction of UMLS is not easy because they are, at least to best of our knowledge, not discussed in the literature. The problems of vocabulary integration into one common system are mostly discussed in
the context of technical solutions and practical issues. The principles of the UMLS construction have not been questioned neither were they related to theories of knowledge influencing our field. In this thesis, we attempted to do a critical evaluation of the UMLS and relate it to the theories of knowledge.

From the historicist and pragmatic perspectives, meaning and concepts are context-dependent and semantic relations are context-relative. From this point of view it is therefore problematic to merge various knowledge organizing systems developed in various contexts for a range of different purposes together, to create one common system that disregards this context-ladeness. The pragmatic perspective has been influential in the approach to knowledge organization called domain analysis. This approach takes as its starting point the view that in every domain there are different voices and paradigms that may have different needs and interests. Moreover, they may develop their own concepts and terminology, competing with the concepts of other perspectives and paradigms in the domain. Therefore, a fruitful basis for knowledge organizing systems is to uncover these voices and paradigms, to identify the meanings and concepts they use and finally, to inform a user about them in knowledge organizing system. In this way, the user himself could decide what perspective is the most useful for his need. This approach therefore calls for discovering theoretical and historical contexts in which concepts developed by examining the literatures of a particular domain.

From this point of view, the UMLS have not fulfilled the potentials and requirement of the most valuable knowledge organizing system because it does not consider these contexts and perspectives in biomedicine. In the way terms and semantic relations are integrated in the UMLS metathesaurus, we may find the traces of positivist and empiricist influence. Mostly, we can see this in the way biomedical vocabularies are integrated together as if they were neutral and objective representations of biomedical knowledge, as if they were reflecting the true nature of reality. The tendency to believe in neutral reality that is beyond human mind, as well as society's influence, has always been stronger in natural sciences than in humanities and biomedical domain is no exception. Even the modern science is still driven by the idea of discovering the real scientific truth, the only right principles and laws governing the phenomena of interest, in the case of biomedical domain, principles that rule various areas of biomedical research.

However, if it is the case that there are different paradigms and perspectives in biomedicine, such as the Evidence-Based Medicine and its perhaps more hermeneutic and historicist oriented counterpart, than by ignoring these perspectives and regarding the concepts and meaning as context-independent and universal we are depriving the users of knowledge organizing systems of potentially the most useful
It may nevertheless be the case that there is more agreement among scientists in biomedicine about the nature of their research and about the concepts used to describe their domain, than there is, for example, in arts or other humanities disciplines in which most of the domain analysis research has been carried out. It is an important function of concepts to stabilize meaning. In biomedicine, without this stability, there would be no agreement on how to cure a particular disease or how to examine a patient with certain condition. And indeed, sometimes there is not such a consensus. But in most cases there are views shared by most researchers and practitioners in the field. It is reasonable to expect that in the biomedical domain the meaning and concepts are more stabilized.

From pragmatic and historicist perspective, the UMLS disregards the importance of context for the development of concepts. However, it would not be right to deny that the UMLS recognizes the importance of some context, i.e. the context of the individual source vocabularies from which it is built. From the very beginning, the project aims at uncovering the differences between the source vocabularies and tries to transparently represent them in the metathesaurus. In each concept entry, it is possible to see the contextual information from every single source vocabulary. The context in the sense of the individual vocabulary context may be restored and retrieved in the metathesaurus. Every concept is in fact a combination of multiple semantic representations in various knowledge organizing systems. This feature makes the UMLS an excellent tool for biomedical vocabulary mapping and translation, as well as an efficient semantic tool for retrieval of biomedical information from a variety of sources. The numerous applications that rely on the use of UMLS indicate that in this respect the UMLS has been a great success. We are not trying to deny this neither decrease the importance of such an achievement for knowledge organization.

To develop a knowledge organizing system of the size and complexity of the UMLS that would be based on the historicist and pragmatic view of concepts, i.e. that would reveal the historical development of each concept as well as various social contexts and scientific theories that might have influenced the understanding of that concept is a theoretically important and pioneering idea which is, unfortunately, unfeasible in practice.

The Unified Medical Language System is in a difficult position because it must find the balance between the variety and context-relativeness of meaning and concepts on the one hand, and the need for biomedical vocabulary unification and integration on the other. This is the biggest challenge
knowledge organization must face. Although the UMLS does not fulfill the requirements of the post-Kuhnian view of concepts, it does offer a certain middle way between vocabulary standardization and meaning-relativeness. It preserves the unique views of meaning provided by the individual source vocabularies while at the same time providing the framework in which they can be integrated and unified.

The important conclusions from the discussion above are that the theories of knowledge play an important role in knowledge organization and often unintentionally and implicitly underlie both theoretical and practical work in our field, and that knowledge organizing systems should be built on principles that are explicit about their theoretical basis. These views are of immense, yet not recognized importance.
7 References

- COOPER, R. (2005). *Classifying Madness: A Philosophical Examination of the Diagnostic and


• SCHUYLER, P. L. et al. (1993). The UMLS Metathesaurus: representing different views of


## 8 Appendix

### 8.1 Semantic Types

<table>
<thead>
<tr>
<th>Entity</th>
<th>Physical Object</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Organism</td>
</tr>
<tr>
<td></td>
<td>Plant</td>
</tr>
<tr>
<td></td>
<td>Alga</td>
</tr>
<tr>
<td></td>
<td>Fungus</td>
</tr>
<tr>
<td></td>
<td>Virus</td>
</tr>
<tr>
<td></td>
<td>Rickettsia or Chlamydia</td>
</tr>
<tr>
<td></td>
<td>Bacterium</td>
</tr>
<tr>
<td></td>
<td>Archaeon</td>
</tr>
<tr>
<td></td>
<td>Animal</td>
</tr>
<tr>
<td></td>
<td>Invertebrate</td>
</tr>
<tr>
<td></td>
<td>Vertebrate</td>
</tr>
<tr>
<td></td>
<td>Amphibian</td>
</tr>
<tr>
<td></td>
<td>Bird</td>
</tr>
<tr>
<td></td>
<td>Fish</td>
</tr>
<tr>
<td></td>
<td>Reptile</td>
</tr>
<tr>
<td></td>
<td>Mammal</td>
</tr>
<tr>
<td></td>
<td>Human</td>
</tr>
<tr>
<td></td>
<td>Anatomical Structure</td>
</tr>
<tr>
<td></td>
<td>Embryonic Structure</td>
</tr>
<tr>
<td></td>
<td>Anatomical Abnormality</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Organic Chemical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nucleic Acid, Nucleoside, or Nucleotide</td>
</tr>
<tr>
<td></td>
<td>Organophosphorus Compound</td>
</tr>
<tr>
<td></td>
<td>Amino Acid, Peptide, or Protein</td>
</tr>
<tr>
<td></td>
<td>Carbohydrate</td>
</tr>
<tr>
<td></td>
<td>Lipid</td>
</tr>
<tr>
<td></td>
<td>Steroid</td>
</tr>
<tr>
<td></td>
<td>Eicosanoid</td>
</tr>
<tr>
<td></td>
<td>Inorganic Chemical</td>
</tr>
<tr>
<td></td>
<td>Element, Ion, or Isotope</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Body Substance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Food</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Conceptual Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Idea or Concept</td>
</tr>
<tr>
<td></td>
<td>Temporal Concept</td>
</tr>
<tr>
<td></td>
<td>Qualitative Concept</td>
</tr>
<tr>
<td></td>
<td>Quantitative Concept</td>
</tr>
<tr>
<td></td>
<td>Functional Concept</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Body System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spatial Concept</td>
</tr>
<tr>
<td></td>
<td>Body Space or Junction</td>
</tr>
<tr>
<td>Congenital Abnormality</td>
<td>Acquired Abnormality</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Fully Formed Anatomical Structure</td>
<td>Body Part, Organ, or Organ Component</td>
</tr>
<tr>
<td>Tissue</td>
<td>Cell</td>
</tr>
<tr>
<td>Cell Component</td>
<td>Gene or Genome</td>
</tr>
<tr>
<td>Manufactured Object</td>
<td>Medical Device</td>
</tr>
<tr>
<td>Research Device</td>
<td>Clinical Drug</td>
</tr>
<tr>
<td>Substance</td>
<td>Chemical</td>
</tr>
<tr>
<td>Chemical Viewed Functionally</td>
<td>Pharmacologic Substance</td>
</tr>
<tr>
<td>Antibiotic</td>
<td>Biomedical or Dental Material</td>
</tr>
<tr>
<td>Biologically Active Substance</td>
<td>Neuroactive Substance or Biogenic Amine</td>
</tr>
<tr>
<td>Hormone</td>
<td>Enzyme</td>
</tr>
<tr>
<td>Vitamin</td>
<td>Immunologic Factor</td>
</tr>
<tr>
<td>Receptor</td>
<td>Indecator, Reagent, or Diagnostic Acid</td>
</tr>
<tr>
<td>Hazardous or Poisonous Substance</td>
<td>Chemical Viewed Structurally</td>
</tr>
<tr>
<td>Body Location or Region</td>
<td>Molecular Sequence</td>
</tr>
<tr>
<td>Nucleotide Sequence</td>
<td>Amino Acid Sequence</td>
</tr>
<tr>
<td>Carbohydrate Sequence</td>
<td>Geographic Area</td>
</tr>
<tr>
<td>Finding</td>
<td>Laboratory or Test Result</td>
</tr>
<tr>
<td>Sign or Symptom</td>
<td>Organism Attribute</td>
</tr>
<tr>
<td>Clinical Attribute</td>
<td>Intellectual Product</td>
</tr>
<tr>
<td>Classification</td>
<td>Regulation or Law</td>
</tr>
<tr>
<td>Language</td>
<td>Occupation or Discipline</td>
</tr>
<tr>
<td>Biomedical Occupation or Discipline</td>
<td>Organization</td>
</tr>
<tr>
<td>Health Care Related Organization</td>
<td>Professional Society</td>
</tr>
<tr>
<td>Self-help or Relief Organization</td>
<td>Group Attribute</td>
</tr>
<tr>
<td>Group</td>
<td>Professional or Occupational Group</td>
</tr>
<tr>
<td>Population Group</td>
<td>Family Group</td>
</tr>
<tr>
<td>Age Group</td>
<td>Patient or Disabled Group</td>
</tr>
</tbody>
</table>

### Event

<table>
<thead>
<tr>
<th>Activity</th>
<th>Environmental Effect of Humans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior</td>
<td>Natural Phenomenon or Process</td>
</tr>
<tr>
<td>Social Behavior</td>
<td>Biologic Function</td>
</tr>
<tr>
<td>Individual Behavior</td>
<td>Physiologic Function</td>
</tr>
<tr>
<td>Daily or Recreational Activity</td>
<td>Organism Function</td>
</tr>
<tr>
<td>Occupational Activity</td>
<td>Mental Process</td>
</tr>
<tr>
<td>Health Care Activity</td>
<td>Organ or Tissue Function</td>
</tr>
<tr>
<td>Laboratory Procedure</td>
<td>Cell Function</td>
</tr>
<tr>
<td>Diagnostic Procedure</td>
<td>Molecular Function</td>
</tr>
<tr>
<td>Therapeutic or Preventive Procedure</td>
<td>Genetic Function</td>
</tr>
<tr>
<td>Research Activity</td>
<td>Pathologic Function</td>
</tr>
<tr>
<td>Molecular Biology Research Technique</td>
<td>Disease or Syndrome</td>
</tr>
<tr>
<td>Governmental or Regulatory Activity</td>
<td>Mental or Behavioral Dysfunction</td>
</tr>
<tr>
<td>Educational Activity</td>
<td>Neoplastic Process</td>
</tr>
<tr>
<td>Machine Activity</td>
<td>Cell or Molecular Dysfunction</td>
</tr>
<tr>
<td>Phenomenon or Process</td>
<td>Experimental Model or Disease</td>
</tr>
<tr>
<td>Human-caused Phenomenon or Process</td>
<td>Injury or Poisoning</td>
</tr>
</tbody>
</table>
### 8.2 Semantic relations

<table>
<thead>
<tr>
<th>isa</th>
<th><a href="continued">associated_with</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>associated_with</td>
<td>physically_related_to</td>
</tr>
<tr>
<td>physically_related_to</td>
<td>part_of</td>
</tr>
<tr>
<td>part_of</td>
<td>consists_of</td>
</tr>
<tr>
<td>consists_of</td>
<td>contains</td>
</tr>
<tr>
<td>contains</td>
<td>connected_to</td>
</tr>
<tr>
<td>connected_to</td>
<td>interconnects</td>
</tr>
<tr>
<td>interconnects</td>
<td>branch_of</td>
</tr>
<tr>
<td>branch_of</td>
<td>tributary_of</td>
</tr>
<tr>
<td>tributary_of</td>
<td>ingredient_of</td>
</tr>
<tr>
<td>ingredient_of</td>
<td>spatially_related_to</td>
</tr>
<tr>
<td>spatially_related_to</td>
<td>location_of</td>
</tr>
<tr>
<td>location_of</td>
<td>adjacent_to</td>
</tr>
<tr>
<td>adjacent_to</td>
<td>surrounds</td>
</tr>
<tr>
<td>surrounds</td>
<td>traverses</td>
</tr>
<tr>
<td>traverses</td>
<td>functionally_related_to</td>
</tr>
<tr>
<td>functionally_related_to</td>
<td>affects</td>
</tr>
<tr>
<td>affects</td>
<td>manages</td>
</tr>
<tr>
<td>manages</td>
<td>treats</td>
</tr>
<tr>
<td>treats</td>
<td>disrupts</td>
</tr>
<tr>
<td>disrupts</td>
<td>complications</td>
</tr>
<tr>
<td>complications</td>
<td>interacts_with</td>
</tr>
<tr>
<td>interacts_with</td>
<td>prevents</td>
</tr>
<tr>
<td>prevents</td>
<td>brings_about</td>
</tr>
<tr>
<td>brings_about</td>
<td>produces</td>
</tr>
<tr>
<td>produces</td>
<td>causes</td>
</tr>
<tr>
<td>causes</td>
<td>performs</td>
</tr>
<tr>
<td>performs</td>
<td><a href="continued">functionally_related_to</a></td>
</tr>
<tr>
<td>functionally_related_to</td>
<td>carries_out</td>
</tr>
<tr>
<td>carries_out</td>
<td>exhibits</td>
</tr>
<tr>
<td>exhibits</td>
<td>practices</td>
</tr>
<tr>
<td>practices</td>
<td>occurs_in</td>
</tr>
<tr>
<td>occurs_in</td>
<td>process_of</td>
</tr>
<tr>
<td>process_of</td>
<td>users</td>
</tr>
<tr>
<td>users</td>
<td>manifestation_of</td>
</tr>
<tr>
<td>manifestation_of</td>
<td>indicates</td>
</tr>
<tr>
<td>indicates</td>
<td>result_of</td>
</tr>
<tr>
<td>result_of</td>
<td>temporally_related_to</td>
</tr>
<tr>
<td>temporally_related_to</td>
<td>co occurs_with</td>
</tr>
<tr>
<td>co occurs_with</td>
<td>precedes</td>
</tr>
<tr>
<td>precedes</td>
<td>conceptually_related_to</td>
</tr>
<tr>
<td>conceptually_related_to</td>
<td>evaluation_of</td>
</tr>
<tr>
<td>evaluation_of</td>
<td>degree_of</td>
</tr>
<tr>
<td>degree_of</td>
<td>analyzes</td>
</tr>
<tr>
<td>analyzes</td>
<td>assesses_effect_of</td>
</tr>
<tr>
<td>assesses_effect_of</td>
<td>measurement_of</td>
</tr>
<tr>
<td>measurement_of</td>
<td>measures</td>
</tr>
<tr>
<td>measures</td>
<td>diagnoses</td>
</tr>
<tr>
<td>diagnoses</td>
<td>property_of</td>
</tr>
<tr>
<td>property_of</td>
<td>derivative_of</td>
</tr>
<tr>
<td>derivative_of</td>
<td>developmental_form_of</td>
</tr>
<tr>
<td>developmental_form_of</td>
<td>method_of</td>
</tr>
<tr>
<td>method_of</td>
<td>conceptual_part_of</td>
</tr>
<tr>
<td>conceptual_part_of</td>
<td>issue_in</td>
</tr>
<tr>
<td>issue_in</td>
<td></td>
</tr>
</tbody>
</table>

### 8.3 Abbreviations

- **CUI**: Concepts Unique Identifier
- **EBM**: Evidence-Based Medicine
- **GO**: Gene Ontology
- **IR**: Information Retrieval
- **IS**: Information Science
- **KO**: Knowledge Organization
- **KOS**: Knowledge Organizing System/Systems
- **MeSH**: Medical Subject Headings
- **NLM**: National Library of Medicine, U. S.
- **NLP**: Natural Language Processing
- **UMLS**: Unified Medical Language System

### 8.4 Figures

**Figure 1** - UMLS Knowledge Source Server: Metathesarus view – p. 14