Information Retrieval and Cognitive Research

Gercina Ângela Borém Lima*, K. S. Raghavan**

*School of Information Science, Federal University of Minas Gerais,
Belo Horizonte, MG, Brazil, E-mail:glima@eci.ufmg.br, home page: www.eci.ufmg.br/glima
**Department of Information Science, University of Madras, Chennai, India,
E-mail: ksragav@hotmail.com; raghavan@unom.ac.in

Gercina Ângela Borém Lima teaches at the School of Information Science, Federal University of Minas Gerais (UFMG), Brazil. She holds a Master's degree in Library Science from the Clark Atlanta University (USA) and a Ph.D. in Information Science from UFMG. She has published papers in periodicals in Brazil, Colombia, Argentina and Portugal. Her current research interests include knowledge organization in the context of digital libraries.

K.S. Raghavan is Dean (Academic) and Professor & Head, Department of Information Science, University of Madras, India. He was a senior Fulbright Scholar in UCLA in 1987. He served as a visiting professor in the School of Information Science, UFMG, Brazil between March and June 2003. He has co-edited several volumes including Digital Libraries: Dynamic landscapes for knowledge creation, access and dissemination: Proceedings of the 4th International Conference of Asian Digital Libraries (ICADL, 2001), Information Services in a networked environment in India (INFLIBNET Centre, 2000), Cognitive Paradigms in Knowledge Organisation: Proceedings of the 2nd International Conference of ISKO (Sarada Ranganathan Endowment for Library Science, 1992) His current research interests include knowledge organization in digital libraries, multiscript databases and informetrics.

Lima, Gercina Ângela Borém and Raghavan, K. S.. (2004). **Information Retrieval and Cognitive Research.** *Knowledge Organization, 31(2).* 98-105. 22 refs.

ABSTRACT: Information Science, which attained the status of a discipline in the 1980s, has been enriched by inputs from a number of disciplines ranging from computer technology to psychology. A predominant characteristic of research in Information Retrieval in recent years has been the adoption of a 'user-centered' approach to the design of information systems. This shift in the emphasis began primarily after Belkin enunciated his ASK (Anomalous State of Knowledge) hypothesis. Research in the Cognitive Sciences has the potential to contribute substantially to enhancing all Information Retrieval processes. This paper emphasizes the importance of adopting a broad-based approach to cognitive research in **IR** and suggests that there is a need for exploring the relevance of analytico-synthetic approach and related research in the design of **IR** systems.

Introduction

Information Science (**IS**) as an independent field of study, came into being in the beginning of the 1960's. There are several definitions of **IS** – each emphasizing some particular facets or components. Some definitions emphasize the aspects of storage, management and dissemination of information; some emphasize the links with technology; and some others the links with information systems (Management Sciences) and communication processes. In effect, the large number of definitions and the different viewpoints merely tend to highlight the interdisciplinary nature of **IS**. In an attempt to illustrate the diversity of perceptions about **IS**, Silva (1999, p. 105) introduced a schema indicating the various view points of experts. The focus of the present paper is to merely bring out the characteristics of **IS** as an independent inter--disciplinary field and to highlight that:

– Information Retrieval (**IR**) is at the core of **IS**; and

– There is a paradigm shift in **IS** with implications for **IR** research

Towards this end three definitions of **IS** offered by Borko, Foskett and Saracevic are quoted below. Borko defined **IS** as below:

> Information science is a discipline that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with that body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information… It has both a pure science component, which inquires into the subject without regard to its application, and an applied science component, which develops services and products. (1968, p.3)

Foskett defined **IS** as:

> A discipline that emerged from a 'cross fertilization' of ideas that include an old art of librarianship, a new art of computer science, arts of new types of communication and those sciences such as psychology and linguistics that, in their modern forms are related directly with all problems of communications. (1980, p.64)

According to Saracevic, one of the most important theoreticians in the field:

> Information Science is a field devoted to scientific inquiry and professional practice addressing the problems of effective communication of knowledge and knowledge records among humans in the context of social, institutional, and/or individual uses of and needs of information. In addressing these problems of particular interest it is taking as much advantage as possible of the modern information technology. (1996, p.47)

While all the three definitions emphasize the interdisciplinary nature of **IS**, the definition given by Saracevic is particularly relevant for the purpose of this discussion as it highlights **IR** not only as the core of **IS** but also as the principal cause for the emergence of the discipline of **IS**. The definition also touches on individuals in the process of communication. Saracevic (1996, p.48) identifies four disciplines closely related to **IS**: Librarianship, Computer Science, Cognitive Sciences and Communication (Fig.1).

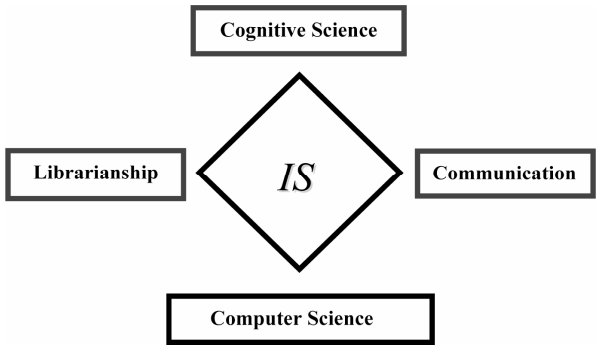

FIG.1 – **IS**

That Librarianship bears a strong relationship to **IS** is no surprise as the two are considered to share very nearly the same objective. Saracevic (1996, p. 50) talks about the differences between Computer Science and **IS**: "... computer science is about algorithms that transform information, and information science is about the very nature of information and its communication for use by humans". There are, on the other hand, common areas of interest including representation and organization of information, **IR** and digital libraries. The definitions of information as a phenomenon and of communication as a process form the basis for linkages between **IS** and Communication Sciences. The beginnings of **IS** can be traced to the knowledge revolution that followed the Second World War. The increasing volume of information being generated in sciences and technology began to demand a higher level of organization of information and also underlined the need for the construction of a theoretical, empirical and practical edifice. This not only resulted in the development of newer tools and techniques for **IR**, but also encouraged theoretical and experimental studies in such areas as the structure of knowledge and its representation, information behaviour of users, human-computer interaction, measures and methods of evaluation of **IR** systems – to mention a few. It should not therefore come as a surprise that this period saw the emergence of the Uniterm Indexing of Mortimer Taube, the thesaurus as a vocabulary control device and also increased research in the application of facet analysis in the area of information organization. Another major development that followed was the emergence of the information industry – a clear demonstration of **IR** as a

commercial enterprise. As an academic discipline **IS** was formally born in 1962 in a meeting at the Georgia Institute of Technology, and was to concern itself with "...attempts to formalize the properties of information by applying information theory, and several other constructs from cognitive science, logic and/or philosophy…" (Saracevic, 1996, p. 46). The 1970's marked the beginning of a paradigm shift in **IS** with the 'user as the focus'. It is during this period that we see increasing emphasis being placed on the importance of cognitive research to **IS** in general and **IR** in particular.

Cognitive Science

CS (referred to also as 'the science of mind') has developed greatly in the last two decades as one of the newest interdisciplinary fields. According to Mey (1982), **CS** is "a contemporary tool with empirical foundations concerned with long standing epistemological questions, especially those related to the nature of knowledge, its components, origins, development and usage". Casti (1989), quoted by Saracevic (1996, p. 51), says that **CS** is an "... amalgam of psychology, philosophy, anthropology, neuro-physiology, computer science and linguistics, organized around the use of computers as a probe for teasing out the secrets of both the brain and the mind". Some of the areas of interest to cognitive scientists have been discussed by philosophers who addressed such questions as the nature of mental representation, the boundaries of the human mind in controlling processes between reason and feelings, etc. Gardner quotes René Descartes as asserting, in the seventeenth century, the core of his rationale: "the human mind is apart from the human body and operates independently from it, constituting a wholly different entity" (Gardner, 1996, p.65). Three hundred years after Descartes, another philosopher, Jerry Fedor, considered a total 'cognitivist', took a step further when he said that it is possible to believe in the existence of mental states and their consequential efficiency without having to believe in the interaction between mind and matter. He believed in a thought language (Fodor, 1975). If the cognitive system involves representations of the symbol kind, these representations should exist somewhere and be manipulated in a certain way. He goes on to say that the language of thinking should be a rich vehicle to be able to execute the many cognitive processes such as perception, thinking, language learning, etc. **CS** was formally born around 1956, after the Theory of Information Symposium held at the Massachusetts Institute of Technology, where the human scientists and communication scientists presented their research findings. Three factors have significantly contributed to the emergence of **CS** as a major area of study. These are:

- The attempts to program computers to perform human tasks;
- Development of information processing psychology with the goal of understanding the internal logic involved in perception, language learning, memory and thinking;
- Development of generative grammar theory and other derivations in linguistics.

The Cognitive Hexagon (Fig.2) indicates the six major disciplines that have made inputs to **CS**. Interestingly enough, some of these disciplines have also made significant inputs to **IS**.

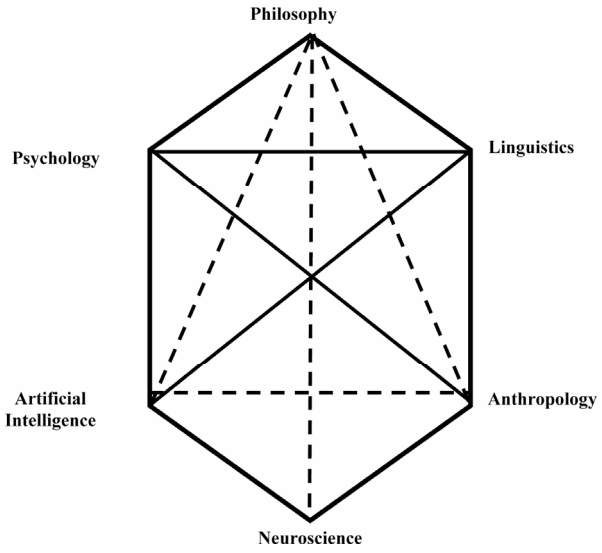

FIG.2 – The Cognitive Hexagon (Gardner, 1996, p.52)

IR and Cognitive Research

There is consensus among researchers in Information Science about the significant inputs that cognitive research can make to the theoretical basis and processes of information representation and retrieval. When computer-based information retrieval systems emerged in the 1960s, they were static batch processing systems. The development of online information retrieval systems enabled interaction between end users

and **IR** systems. The advent of computers in information retrieval triggered two streams of research in **IR**; the first concentrated on using computers to automatically derive document representations and tools for knowledge representation. The second concentrated on using the technology to provide online help to the end user to support more effective interaction between the system and end users during the course of a search. However, in recent years 'interaction' has come to be viewed as an important factor in the process of information retrieval. Not only this, 'interaction' has come to be interpreted to mean much more than user feedback. It has even been suggested that research in **IR** should concentrate more on 'interaction' and apply this knowledge in the design and development of more effective **IR** systems in general and better end user interfaces in particular (Saracevic, 1996a). This increasing emphasis on 'interaction' in **IR** probably owes its origin to Belkin's enunciation of the ASK hypothesis and the shift in **IS** to a user-centered paradigm. An examination of the relevant literature suggests that the principal issues discussed in the literature in this area relate mainly to users' prior knowledge, and the cognitive processes involved in the interaction with the system including the act of processing information. It does indeed appear that by and large cognitive research in **IR** is largely restricted to ways and means of enhancing the end-user interaction with the system during the search process. It is important to take a broader view of **IR.** It is emphasized here that all the processes involved in **IR,** including subject analysis and representation are activities involving information processing and are, therefore, essentially cognitive in nature. The argument made here is that it is important to take such a 'holistic' view for enhancing and for realizing a qualitative improvement in **IR**.

The cognitive approach starts with the assumption that all information processing activities are interactive and are mediated by a system of categories that represent the world model of the information processor. The cognitive viewpoint in **IR** implies that every information processing activity is a cognitive process influenced by the cognitive structure or world model of the information processor. The **IR** process revolves around a number of cognitive players. As Ingwersen (1996, p.11) puts it: "when seen from a cognitive perspective all of interactive communication activities in **IR** and information seeking can result in processes of cognition which may occur in all the information processing components involved". Brookes (1977) was one the first authors to use the cognitive point of view in an attempt to develop a theory of IS. In order to understand the inputs that cognitive research can make to enhance **IR**, it is important to begin by delineating the processes involved in **IR**. However, much of cognitive research in **IR** appears to focus on the end-user as a processor of information received from an **IR** system that stores surrogates of information objects such as documents. The fact that the system acts as an information processor based on the information it receives from the user of the system and on the information space that has already been created by the **IR** system developers is something that is not widely discussed. The processes of creating document surrogates (i.e., the process of indexing) and search formulation are as much cognitive processes whether carried out by humans or machines. The cognitive structure of the surrogates of information objects in an **IR** system at any given point in time, therefore, is what it is – a product and function of the cognitive structures of all those components that created it. The surrogates created for an information object, e.g., index records and/or abstract of a document or a search expression are products of information processing activities and involve the cognitive structures and world models of the information object being processed and the indexer/abstractor/user/searcher. The cognitive structure of the information object, say a print document – that is being processed by the indexer/abstractor is static. The effectiveness of the surrogates in adequately and accurately representing the cognitive space of the information object is therefore largely a function of:

- The conceptual knowledge of the information processor (i.e., the indexer /abstractor);
- The subject matter and the structure of the text being processed;
- The language used for representing and the information processor's knowledge of this;
- The information processor's understanding of the users of the system; for example, it is important for the information processor to see all the viewpoints or aspects of the information object being processed from the perspectives of potential end users of the information object.

What goes on in an **IR** system? The primary objective of an information retrieval system is to bring together information resources and end users of information who may benefit from these at the time of searching for information. The process of information generation by authors of texts, although a cognitive process,

is something that has already taken place over which the **IR** system has no control. In fact, information generation is one of the causes of **IR**. However, the surrogates that are intended to represent information objects (e.g., texts) are expected to semantically reflect the world models of authors of texts. Basically, there are two broad types of operations involved in **IR**. The first involves representation or mapping the knowledge/information in a documentary resource and/or mapping the information requirements of a user (both used as inputs in an **IR** system). The second type involves enabling and supporting the process of reducing the semantic gap between the user and the system through a process of interaction between the user and the system. An adaptation of the **IR** model developed by Ingwersen and presented in the following figure brings out the cognitive processes involved in such a communication system.

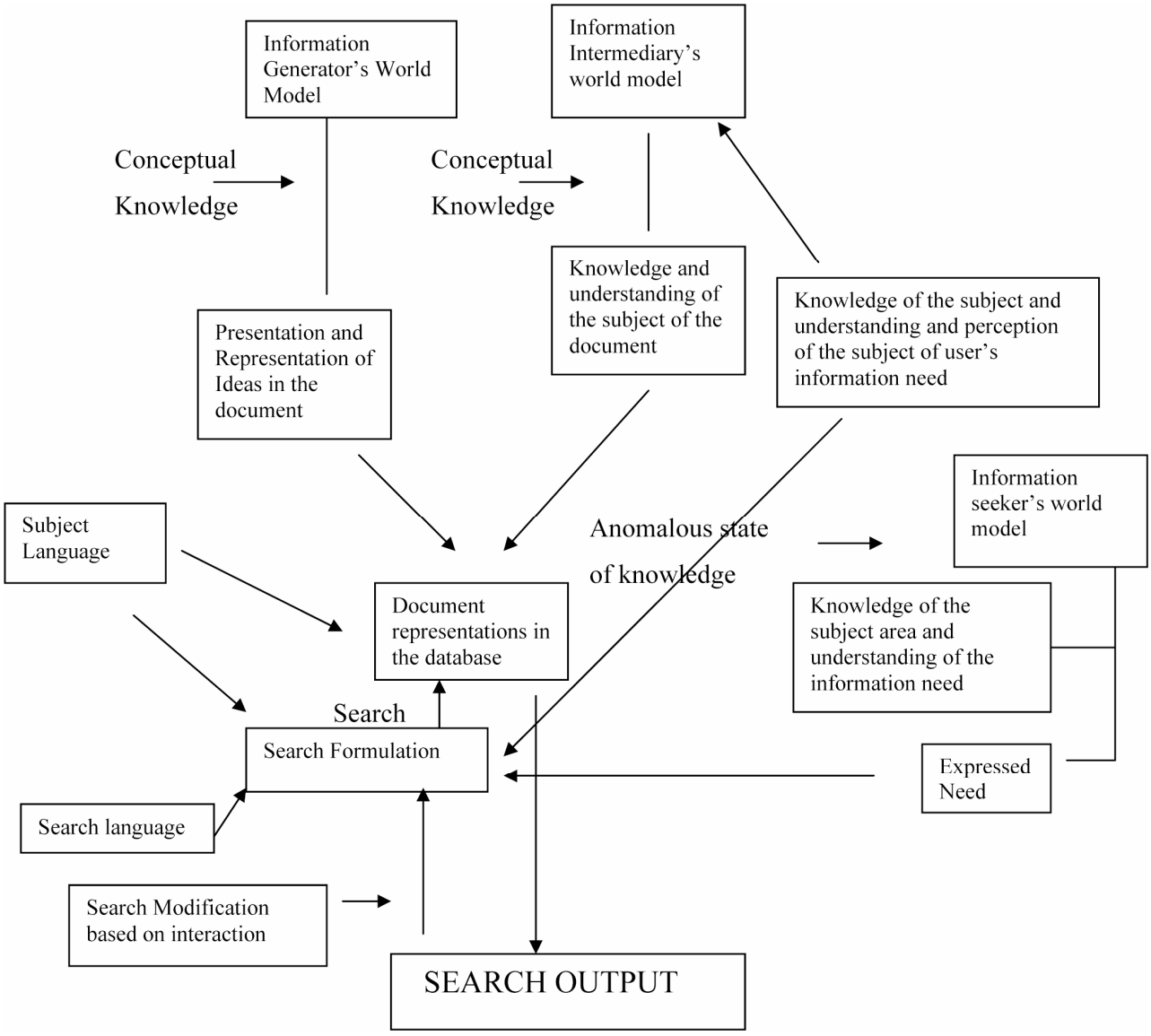

*Fig. 3 **IR** as a cognitive process*

Ingwersen's model is in effect a representation of the total formal information transfer cycle. The information processing activities of all the three major players in the total information transfer process, viz., the generator of information, the user of information and the information system, are coloured and affected by their respective cognitive structures or world models. For example, the conceptual knowledge as presented in a document will necessarily be coloured by the author's world model. Similarly the processes of index-

ing, query formulation, search formulation, etc. will be coloured and affected by the world models of the respective players. Once we concede this, it should not come as a surprise that evaluation studies of the 1950s and later found individuals not agreeing on relevance judgments. From a cognitive viewpoint the central problem for research in **IR** is: 'How to match the cognitive structures of the different players involved in Information Retrieval?'. In other words it is safe to assume that effectiveness of information retrieval could be enhanced if the user's world model corresponds to or can be made to correspond to the world model of the **IR** system. Belkin (1990, p.11) affirms that "…beliefs and so on of human beings (or information-processing devices) mediate (or interact with) that which they receive/perceive or produce". Looked at from such a perspective it is easy to understand why some writers consider **IS** as part of **CS**. (Garcia Marco and Esteban Navarro, 1993, p.11). Both the fields are interested in how information is processed and how it is better adapted to reality.

IR systems have no control over the processes of information generation, but are a result of such processes. Let us consider the following activities involved in **IR**.

- The processes involved in creating surrogates for information objects (subject analysis and representation)
- The search process
- The interface between the system and the user

Some of these processes in **IR** are represented in fig. 3.

An understanding of all these information processing activities will contribute to the design of better **IR** systems. It is, therefore, important to examine if there are universals in the system of categories that mediate information processing activities. The efforts should focus on building a schema that approximates an information processor's "minimal mental model". Marc de Mey's hypothesis that "any ... information processing, whether perceptual (such as perceiving an object) or symbolic (such as understanding a sentence) is mediated by a system of categories and concepts which, for the information processor constitutes a representation or a model of his world". (Mey, 1982, p.4) is important in this context. Search for universals in **IR** research is not something new. The notion of 'Absolute Syntax' postulated by S. R. Ranganathan is based on such an assumption. Simply defined, Absolute Syntax refers to a subject representation that maps the ideas/concepts in a manner that is closely parallel to how these ideas would be mapped in the mind of a human information processor (generator or user) in his or her thought process. Neelameghan (1979) found parallels in a variety of disciplines to support the notion that such a model exists. A series of experimental studies on multilingual aspects of PRECIS suggest that knowledge representation could be largely independent of linguistic syntax confirming again that universals could exist (Austin, 1976, Sorensen & Austin, 1976, Lambert, 1976). Jacob and Shaw (1998, p.155) say that "categorizing is a cognitive process of dividing the world of experience into groups of entities, or categories, to construct order out of the physical and social world(s) in

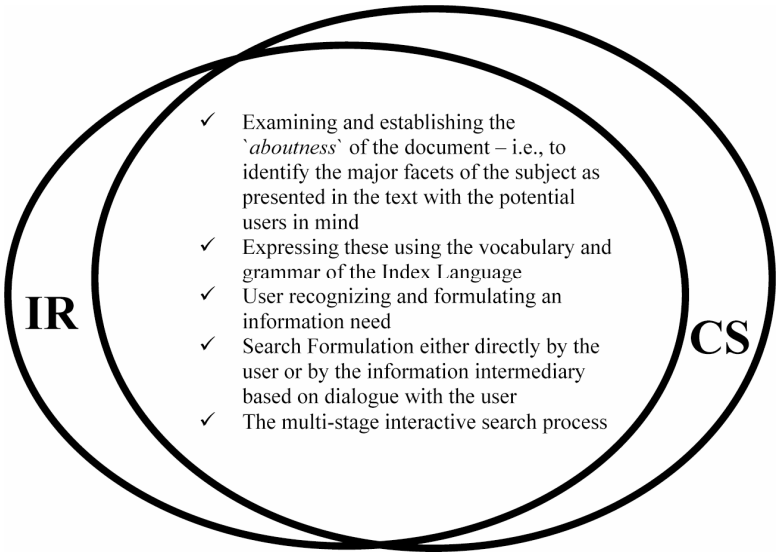

Fig.3 –IR and CS

which the individual participates". Markman (1989), quoted by Jacob and Shaw (1998, p.155) describes it as "a fundamental cognitive mechanism that simplifies the individual's interaction with the environment: it not only facilitates the efficient storage and retrieval of information, but also reduces demands on human memory". For Piedade (1983), this is a common mental process of man, since we live classifying ideas and things automatically, in order to know and understand. Finally, Gardner (1996, p.373) states that "the categories have an internal structure centered on prototypes and stereotypes and, other constituting elements are defined as more or less peripheral depending upon the degree of crucial traits they share with the central prototype". A system of universals should therefore be extremely valuable not only in structuring information to formulate their meaningful representations, but also in facilitating browsing and navigation for effective interaction between the user and the system in an effort to reduce semantic gaps between their respective world models.

Foskett argues that human beings impose structure on their world by organizing and categorizing. A pre-requisite for this is the process of analysis. The two processes together – analysis and synthesis – have been the underlying themes of the Analytico-Synthetic approach propounded by S. R. Ranganathan. However, IR research today appears to focus more on using information technology for processing of texts to derive surrogates than on understanding the more fundamental issues of the nature of information processing itself. It is proposed here that since there is enough evidence to suggest that all information processing is analytico-synthetic in nature and is mediated by a schema of categories representing the world model of the information processor. There is therefore a need for IR research to focus on understanding the common principles that underpin all information processing – be it presentation of ideas in a discourse, indexing/classifying the subject of a document, formulation of a query, analysis of user's need or formulating a search expression. This necessarily has to be done by adopting a multi--disciplinary approach to identify commonality in such information processing activities as learning, information seeking, systems analysis and design, knowledge organisation by authors in presentation of texts, classification and indexing and decision-making, to mention a few. For example, in the context of providing for better and more effective interaction between end users and information systems, researchers are concentrating on Artificial Intelligence (AI) and Human-Computer Interaction (HCI) as two of the most important areas.

AI aims at reproducing the mental activity in such information processing tasks as learning and comprehension. Indexing and formulating a search expression similarly involve mapping cognitive structures. The common denominator that governs all these activities is the 'analytico-synthetic' approach. If indeed all information processing activities in different contexts such as those mentioned above are closely parallel and are characterized by similar modes of thinking, it is highly probable that processes associated with IR could be significantly enhanced by a better understanding of the factors that govern these processes. Research should focus on identifying the commonality in information processing activities in a wide variety of contexts and seek to employ this knowledge in the design of information systems. The notion of 'Absolute Syntax' postulated by S. R. Ranganathan (1967, Sec. 7) and similar notions on universals have relevance for all information processing activities in IR including information analysis and representation, design of end user interfaces, etc. as all these are cognitive processes governed by similar thought processes. For example, there is not much research on the possible use and application of facet analysis and a categorical approach in providing for navigation (hyperlinking) between concepts in designing end-user interfaces. The ongoing research project entitled **Facet** at the University of Glamorgan, U.K. is investigating methods of integrating the thesaurus into the user interface, including the design of a query editor to facilitate construction of multi-term faceted queries (2004).

Conclusion

The field of **IS** has been consolidating its viewpoints since its recognition as a scientific discipline. This process has strengthened in recent years as a direct consequence of the strong links the discipline has established with technology, cognitive sciences, etc. In recent decades there has been greater emphasis in approximating the **IR** system to the mental model of the information processor. While much of the research in cognitive information retrieval has concentrated on the end user, it is proposed here that cognitive research has implications for all IR processes. There is need for research to explore the relevance of developments in the theory and practice of knowledge organization and information representation towards this end.

References

Austin, Derek. (1976). PRECIS in a multilingual context: Part 1– PRECIS: An overview. Libri 26(1), 1-37.

Belkin, Nicholas J. (1990). The cognitive viewpoint in information science. Journal of Information Science 16, 11-15.

Borko, H. (1968). Information science: what is it? American Documentation, v.19(1), 3-5.

Brookes, B.C. (1977). The developing cognitive viewpoint in information science. (In: CC-77: International Workshop on the Cognitive Viewpoint. Ghent: Ghent University. p.195-203).

The Facet project (viewed on Aug. 09, 2004) (Available at http:\\www.comp.glam.ac.uk/~FACET).

Fodor, Jerry A. (1975). The language of thought, New York: Thomas Crowell.

Foskett, D.J. (1980). A Ciência da informação como disciplina emergente: implicações educacionais. Ciência da Informação ou Informática, 53-69. Rio de Janeiro: Calunga.

Foskett, D.J. (1992). Ranganathan and 'User-Friendliness'". Libri 42, 227-34.

Garcia Marco, Francisco Javier and Esteban Navarro, Miguel Angel. (1993). On some contributions of the cognitive sciences and epistemology to a theory of classification. Knowledge Organization. 20(3), 126- 132.

Gardner, Howard (1996). A nova ciência da mente: uma história da revolução cognitiva. Sâo Paulo: EDUSP.

Ingwersen, Peter (1996). Cognitive perspectives of information retrieval interactions: elements of a cognitive IR theory. Journal of Documentation 52(1), 3-50.

Jacob, Elin K., and Shaw, Debora. (1998). Socio-cognitive perspectives on representation. Annual Review of Information Science and technology, 33, 1998, 131-185).

Lambert, Germaine. (1976). PRECIS in a multilingual context: Part 4 – The application of PRECIS in French. Libri 26(4):302-324.

Mey, Marc De. (1982). The cognitive paradigm: an integrated understanding of scientific development. Chicago: The University of Chicago. 314 p.

Neelameghan, A. (1975). Absolute syntax and the structure of an indexing and switching language (In Ordering systems for global information networks. Bangalore: Sarada Ranganathan Endowment for Library Science and FID C/R, 1979).

Piedade, M.A. (1983). Requião introdução à teoria da classificação. Rio de Janeiro: Interciência.

Ranganathan, S. R. (1967). Hidden roots of classification. Information Storage and Retrieval 3, Sec. 7.

Saracevic, Tefko. (1996) Ciência da informação: origem, evolução e relações. Perspectivas em Ciencia da Informaçao 1(1), 41-62.

Saracevic, Tefko. (1996a) Modeling interaction in information retrieval (IR): a review and proposal. (In Proceedings of the American Society for Information Science, 33; 3-9).

Silva, Júnia Guimarães (1999). Ciência da informação: uma ciência do paradigma emergente. (In: Pinheiro, Lena Vania R. eds. Ciência da informação, ciências sociais e interdisciplinaridade, 79-117. Brasília: IBICT)

Sorensen, Jutta, and Derek Austin.(1976). PRECIS in a multilingual context: Part 2 – A linguistic and logical explanation of the syntax. Libri 26 (2):108-139.

Sorensen, Jutta, and Derek Austin. (1976). PRECIS in a multilingual context: Part 3 – Multilingual experiments, proposed codes, and procedures for the Germanic languages. Libri 26 (3):181-215.