Abstract: This paper focuses on the need for knowledge organization (KO) tools, such as library classifications, thesauri and subject heading systems, to be fully disclosed and available in the open network environment. The authors look at the place and value of traditional library knowledge organization tools in relation to the technical environment and expectations of the Semantic Web. Future requirements in this context are explored, stressing the need for KO systems to support semantic interoperability. In order to be fully shareable KO tools need to be reframed and reshaped in terms of conceptual and data models. The authors suggest that some useful approaches to this already exist in methodological and technical developments within the fields of ontology modelling and lexicographic and terminological data interchange.

1. Semantic interoperability, the WWW and knowledge representation (KR)

The evolution of computer networks has stressed the need for data and information interoperability, i.e., for having data and information assets reusable across distributed and heterogeneous systems. According to Amit Sheth (Sheth, 1999), we are now in the third generation of interoperable systems where the concerns are mostly focused on information and knowledge, emphasizing semantic interoperability at a level higher than that of previous developments. Before the expansion of the Internet, interoperability was concerned mainly with intersystems communication and agreements on data syntax and structure for communities of systems (multidatabases, federated databases or federated systems). The WWW brought a new dimension to fundamental concepts such as distribution – from the enterprise-wide space to the global space – and heterogeneity, implying changes in systems paradigms that are far more complex than just a matter of scale. Another fundamental aspect that also became more complex to deal with is autonomy – a systems requirement that has to be balanced with the increased demands in network interoperability. Such demands have influenced turning points in systems’ architecture and design, highlighting the trend for ‘composability’ of solutions, in which components tend to be system independent, adaptable, extendable and reusable. This is true both for software engineering and information design and data modelling.

The trend described by Sheth – from system, syntax, structure to semantics – is well illustrated by the semantic Web movement, and all developments around it. This is especially true of XML, as an independent language to structure resources, and of RDF, as an XML specification to convey machine-understandable
representations of resource descriptions, including content description, and metadata modelling, to build such representations. On top of that, the use of common or shared formal languages (i.e. controlled vocabularies) to convey explicit and shareable representations, as well as of ontologies to support them, became part of what is now understood as the architecture for the future Web (Berners-Lee, Hendler and Lassila, 2001).

Developments around formal languages, ontologies and vocabularies touch three major fields: the field of computing and IT, mostly devoted to raising methods and methodological tools for building controlled knowledge representations (KR), the field of terminology and lexicography, and the field of knowledge organization, where classification systems, thesauri and other controlled vocabularies for information retrieval are largely produced. Most of the theoretical and methodological research around ontologies has flourished and intensified in the last decade to semantically support expert systems (Sheth, 1999; Kashyap and Sheth, 2000). It encompasses diversified levels of approach, from the theory of knowledge representation (Sowa, 2000), to applied fields such as knowledge bases, new methods of software engineering (Gruber, 1993; Guarino, 1998, Guarino and Welty, 2000), or information brokering based on metadata for knowledge domains (Kashyap and Sheth, 2000). Because the use of ontologies in these approaches always imply some form of formalized logic, they developed mostly from the theoretical background of artificial intelligence (AI). Nevertheless, the methods and tools to formalize and structure such ontologies have evolved in ways that are becoming more amenable to use by non AI experts. Therefore, they provide opportunity for exploring new principles and solutions that can benefit renewal in related areas such as knowledge organization for information retrieval.

2. Disclosing knowledge organization (KO) tools in the network

Not only ontologies have emerged “from academic obscurity into mainstream business and practice on the Web”, as noted by McGuinness (McGuinness, 2001), but also the Web environment has raised the importance and value of existing KO tools, such as classifications, thesauri, taxonomies, subject-heading systems etc. These are now claimed to have an enormous potential, not only in being applied to the self description of individual Web documents but also in supporting search and retrieval services provided by agencies other than libraries. This is evident from both the current literature on the matter and from the number of projects, agencies and metadata schemes, etc. that recommend or refer to the use of traditional library KO tools. Apparently, these are ready to use and well credited because they are professionally prepared resources and because they reflect literary warrant. Nevertheless, they show actual constraints that are not minor drawbacks to the objectives of KO tools being widely shared in the network. Such constraints are analysed from several major perspectives in the sections following.

2.1 Explicit ontological frameworks

Although KO tools represent intellectual constructions at its best, backed by conceptual principles, these are often not explicitly conveyed, i.e., in forms that could speak for the system as a whole. This addresses information to clearly identify a system’s domain, its boundaries, structure and major changes during the course of its evolution, categories of concepts and principles governing their relationships, policies regarding relationships to other KO systems, etc. In fact,
what is usually available in the first plan, even for the professional user, is the resulting product - the structured vocabulary of the thesaurus, the subject heading list, the classification schedule - not the underlying philosophy, principles and policies, which in many cases happen to be documented long after the corpus exists. In the larger context of standards, the few international guidelines available in the field (such as ISO 2788 and 5964, for the construction of monolingual and multilingual thesauri, respectively) have long been recognized as insufficient. They are very basic, about three decades old, do not provide for every kind of KO system and do not have the strength to create an agreed ontology for the universe of discourse in the KO field. It is even a paradox that the field of knowledge organization still lacks a sound basis for conceptual and terminological consensus, as it is true that actual tools are still biased by their historical and local technical traditions.

2.2. Open network availability

As resources in their own right, i.e. apart from being provided as discrete elements in bibliographic systems, KO tools only recently became available on the network, although since the nineties some of them have been published in electronic form, e.g., in CDROM as an end-product or just as structured electronic files to be handled by a database system. Besides being available on the network, KO tools need to be designed for different users and usages, including the possibility of being accessed and used by automated agents. For instance, it is important to consider KO tools as identified and maintained persistent namespaces, and to include a structure to allow external entities to link and refer to any of their constituent elements. This is a requisite that fits into the framework of the Web semantic architecture and that may in turn result in modelling requirements at the level of data model or systems architecture.

2.3. Data sources shareability

Data sources shareability refers to the level of portability of data content among different systems, for the purpose of reuse. Several levels of pre-requisites are needed for this, which will be presented starting at the lowest level. Data representation language is the basic level and should not be particular of a given system, group of systems, or platform. This naturally points to XML, if not as the language for holding data in a system, then at least as the possibility of transforming data into XML at the export level. The next level is data structure - i.e., the data representation aspects that determine levels of syntactical compatibility (machine-readable aspects of data, such as data element definition, also referred to as formatting) and logical compatibility, referring to the main components of the underlying model (type and definition of entities, attributes, relations). Data structure should provide for data interchangeability among systems deemed relevant to each other, at least at a minimum level. In practical terms, this would make data transportable without loss of significant content and features, and ensure that it is reusable without major conceptual conflict. These aspects require some level of conformity to common representation standards and declarative tools about data components, features and options particular to each system, at a meta-meta level. With regard to data sources shareability, KO tools used and/or maintained by libraries have improved little by the efforts of library automation. So far, the automation of KO tools assumes one of two forms: either an authority MARC file, designed mainly as a subsidiary management resource internal to a given bibliographic system, or an independent database, primarily oriented for editorial
and publishing (until recently, mostly printed publishing) objectives.

In the first case – MARC authority files, either for subject alphabetical or classification systems - the underlying functional and data models derived mostly from bibliographic management requirements and did not evolve according to other purposes that could be additionally envisaged for the KO tool as an independent resource. For example, MARC authority files have not been used to display the KO tool itself independently of bibliographic data. While in this case the solutions provide for data shareability within the community of MARC systems, such data structure is difficult to share with non-MARC systems, as it is. Simply raising a non-MARC format from scratch would not solve the problem, because the primary community of KO tools are libraries and they all talk MARC, so MARC formats will continue to evolve. One example is the UNIMARC for Classification Format, recently issued by IFLA, whose objectives are twofold: to support authority control functions in bibliographic systems and to serve as a common format used by publishers to deliver classification data. Additionally, it can also serve as the logical format used in classification management systems. Nevertheless, one has to recognize that the model underlying MARC for classification did not explore what were the additional requirements for that purpose.

In the second case – KO systems managed independently of library systems – there are no standards or common data models whatsoever. This is the case with most solutions used to manage thesauri and classification schemes, meaning two things: first, that the shareability of such data in a machine-readable form is not even part of the requirements of such systems; and second, that efforts in the modelling of such KO tools and supporting systems remain isolated and idiosyncratic. They do not aggregate the modelling needs of a community and their data structures remain mostly invisible. Yet, in cases where electronic files are provided, they usually imply a replication of the solution used by the provider, and the knowledge particular to its format.

2.4. Data content usability

There are two different aspects regarding data content usability. First, the clarity of intended meaning of data element instances. This refers to the level of semantic expressiveness provided, by means of definitions, attributes and relations included in the actual content of KO tools, and also to the way they are conveyed to the user. Second, the aspect of alternative usages, i.e., different views of the same data content for different kinds of users and in contexts other than those where the tool originated or to which it was primarily designed. In both respects more could be achieved in terms of usability, if the functional requirements were revised on the assumption that the tool is to serve a wider community of users who are not only end-users – in the sense of bibliographic searchers, or information seekers in general - but also metadata information providers. On one hand, KO tools could consider expanding content by interlinking with other network resources, e.g. referring to other KO tools, or to major resources about a given concept, or instances of a class. On the other hand, data presentation models could be reanalysed in order to improve methods of content and context display, for example by overcoming the traditional use of "condensed" conventions (like symbols and abbreviations) for which the major justification in the past was the "economy of space" imposed by manual products. Today’s systems have no reason to live by old assumptions and limitations, if the target usages are changing and if the technical means available make changes affordable.
2.5. The centrality of data modelling

From all the above perspectives one can easily derive many aspects that would not only modify the model of the traditional KO tools, but could also lead to an assessment of data models. Data models, and standards that support them, are critical for shareability as they address the real technical and semantic conditions without which network availability and content usability are of limited use. The centrality of data modelling comes from its nuclear position in the design chain: it conveys major elements of the conceptual model (elements upon which information is designed), it interprets requirements and it provides support for functionalities. The definition of data standards is the most visible result of data modelling, especially in the case of a new standard. But the overall implied activity goes beyond that, it is not merely a stage in development, it is a level of management that should be continuously exercised, because systems evolve with changes in information. The emphasis on data models is therefore justified, as a critical piece of the information architecture either in individual systems, multi-systems or open information space design. KO managers should begin to think about modelling at large, i.e. expanding the data model beyond the local, individual or specific community's system requirements, either in informational or functional terms, in order to enhance their network value.

3. KO modelling for the future

The conceptual framework for KO tools has much in common with the conceptual framework of an ontology developed for knowledge-based systems. Languages for modelling and representing ontologies have been developed for more than two decades, but not until the advent of the Web was their full potential in information exchange and communication revealed. Also, Web technology itself has helped to advance ontology-modelling languages, through open standards such as XML and RDF.

Before the web, ontology modelling languages were normally characterised by high computational complexity and logic-based formats with specific syntax and a function of mapping ontology from and to computer languages. A typical example is KIF - *Knowledge Interchange Language*. The new generation of Web-oriented ontology languages are now being developed. DAML - *The DARPA Agent Markup Languages*, OIL - *Ontology Inference Layer* and SHOE - *Simple HTML Ontology Extension* are examples of such languages aiming to achieve the status of standards. They maintain the expressive power of a logic-based language while also being system independent and Web compatible through their adherence to XML and RDF. Even languages like UML (Unified Modelling Language), which is primarily a general modelling language for object-oriented analysis and design, is successfully applied in ontology modelling when used with RDF syntax (Cranfield, 2001). The fact that tools for ontology modelling are now made available and understandable even for those without programming and AI expertise opens up new opportunities for the application of ontologies (Ying, 2000). The framework for data analysis and conceptual modelling that is offered by these languages is an important reference point for everyone involved in KO data modelling, as they offer methods for strict analysis of data content, data logic, class definition, property inheritance, etc.

Besides ontological modelling languages, another area of web-based
standards that is of particular interest to the KO field is that of developments supporting terminological exchange. These can also become a source of useful approaches in modelling-for-exchange. Internet multilingual demands have encouraged numerous developments in automatic translation and stimulated research in the field of lexicography and multilingual terminological exchange. Among the first standards for terminology exchange was MARTIF- **Machine Readable Terminology Interchange Format** (ISO 12200), designed as an SGML based format for human-oriented terminological and lexical databases. More recent is OLIF - **Open Lexicon Interchange Format**, an XML-based standard that builds on MARTIF, especially improving the aspect of machine readability of data for machine translation. Building on these two standards, which are primarily focused on lexical data, a new standard is being developed: TMF - **Terminological Markup Framework** (ISO 16642). This one is more networks oriented and includes features for conceptual and ontological aspects of terminological data (Romary, 2001).

These developments from the areas of AI and lexicography could offer a useful background for the reshaping of knowledge organization tools in terms of modelling activities. The aforementioned standards and terminological formatting languages could offer more logical and structured methods for representing linguistic aspects of KO tools in order to enhance them. An initiative that appears to be in line with this trend is Voc-ML - **Vocabulary Markup Language**, raised in the context of NKOS - **Networked Knowledge Organization Systems** promising tsupport to KO tool exchange. However, it is evident that more joint effort from the KO community is required to make advances toward more interoperable and shareable KO models.

### 5. Conclusions

This paper analyses the current state of knowledge organization (KO) tools such as classification schemes, thesauri and authority files, focusing on the strengths and weaknesses of their electronic handling and exploitation in the Web environment. The importance of semantic interoperability stresses the potential of KO information assets in the framework of the Semantic Web, calling for improvement and change in the network shareability of KO tools. This highlights the lack of common methods and standards to support the current needs of cross-domain and multi-system usability of KO data. Existing constraints in the way KO systems are currently managed and made available indicate the need for their significant improvement at the level of modelling, especially in what concerns the modelling of data structures, as a fundamental condition to disclose KO resources over the network. In this respect, an interdisciplinary approach is considered of utmost interest, as is also the case with developments taking place in related areas such as knowledge representation in artificial intelligence and lexicography. These have already raised methods and tools that are not only more aligned with Web technology but also deal with matters and content that are very close to those of knowledge organization.

### References


Open Lexicon Interchange Format (OLIF) [online]. Available at: http://www.olif.net


*All web addresses in the bibliography are checked on 15 March 2002.*