STANDARDS REQUIREMENTS, SOURCES AND FEASIBILITY IN KNOWLEDGE ACQUISITION

B R Gaines[1] & J H Boose[2]

[1]Knowledge Science Institute, University of Calgary
Calgary, Alberta, Canada T2N 1N4

[2]Knowledge Systems Laboratory, Boeing Advanced Technology Center, 7L-64
Boeing Computer Services, P.O. Box 24346, Seattle, WA 98124

Abstract: Standards for the interchange of knowledge structures are becoming increasingly important for the knowledge acquisition research and development community. The sharing of datasets, knowledge bases and tools is now a major activity being promoted through the knowledge acquisition workshops. It is seen as an important stage in the development of the research and associated systems that they are subject to independent use, comparison and evaluation. There is also a major trend towards second generation systems integrating a diversity of data sources, knowledge structures, methodologies, techniques and tools. Such systems involve the use of material from a multiplicity of sources not specifically associated with knowledge acquisition, and their development would be accelerated by the adoption of data and knowledge interchange standards. This paper summarizes the major standards issues for the knowledge acquisition community.

INTRODUCTION

There have now been ten Knowledge Acquisition for Knowledge-Based Systems Workshops, commencing with the first AAAI KAW in Banff in November 1986 which has become the annual North American Workshop, followed by the first European KAW in London, England in June 1987, which has become the annual European Workshop, and extended in 1990 with the first Japanese KAW held in Kyoto in October which is intended to become an annual meeting alternating between Japan and Australia.

Data interchange issues have become major topics of interest to the community that has formed around these meetings, and for the past three years at the North American workshop an auxiliary meeting has been held to attempt to establish standards for interchange of knowledge acquisition data. At the European KAW last year in Amsterdam *Project Sisyphus* was commenced as a data sharing exercise designed to encourage those with different tools related to the same problem types to demonstrate them applied to commonly available datasets. EKAW in Scotland this May has a major part of its program and proceedings devoted to comparative studies. While Project Sisyphus is not primarily oriented to standards issues these are forming a significant part of its management as increasingly complex datasets are submitted to its database.

As Co-Chairs of the Knowledge Acquisition Workshops, we would like to establish strong communications between those considering standards issues for the knowledge acquisition community and those considering them for expert systems, knowledge representation, text processing, hypermedia and so on. In particular, we wish to raise the awareness within the knowledge acquisition community of standards issues and initiatives as they develop.

The knowledge acquisition community has found it necessary to interpret knowledge and its representation rather more liberally than in conventional artificial intelligence studies. Protocol collection, interviewing and other related elicitation methodologies generate large data sets of audio, visual and textual material. Managing this, and the various stages of its analysis, is a major problem in all realistic knowledge acquisition situations. In particular, one wishes to maintain an effective audit trail of links back to this source data as part of the validation, verification, explanation and maintenance processes.

A strong case has also been made, and is part of the architecture of many knowledge acquisition tools, for a variety of mediating representations between the acquisition of expertise and its final computational emulation in formal knowledge structures. There is also growing concern in the knowledge acquisition community with the need to create and maintain knowledge bases in portable form that is relatively independent of the knowledge representation languages of particular expert system shells.

Despite the importance of data interchange and standards issues in knowledge acquisition, it is not seen as appropriate for the knowledge acquisition community to attempt to establish standards that are specific to knowledge acquisition. While the total range of types of material and representation that it is necessary to manage in the knowledge acquisition process may be unique, most of the material can be held in forms that are derivative of the base technologies in use. This is obvious for audio-visual and textual material at one extreme, and for formal knowledge representation at the other. There are some mediating representations, such as the repertory grid, that are peculiar to the knowledge acquisition community, but even these can be reasonably regarded as particular datasets and formats within databases and knowledge bases.

However, it is very appropriate for the knowledge acquisition community to contribute to standards initiatives in a number of ways:

1 By detailing the full range of dataset types involved in knowledge acquisition.
2 By abstracting from current practice a reasonable set of standards requirements that would satisfy most knowledge acquisition needs.
3 By encouraging the application of existent and emerging standards within the knowledge acquisition community.
4 By testing the applicability of proposed standards to knowledge acquisition methodologies.

THE DIVERSITY OF DATA SOURCES AND TYPES IN KNOWLEDGE ACQUISITION

The knowledge engineering bottleneck impeding the effective application of knowledge-based systems, knowledge acquisition has been a topic attracting major research effort and resulting in the development of a wide range of methodologies, techniques and tools (Boose & Gaines, 1988, 1990; Boose, 1989). Early attempts to support the knowledge acquisition process directly focused on facilitating human access to formal knowledge representation languages through improved interfaces, particularly those designed for direct use by domain experts. In parallel with this, as methodologies for interviewing, protocol collection, and mediating representation developed, computer-based tools supporting the storage, indexing, retrieval and processing of large collections of heterogeneous data were developed using hypermedia tools.

Initially the unstructured, informal knowledge was regarded only as a first step towards the creation of formal, computational knowledge structures. However, it has become increasingly apparent that the way in which the formal knowledge derives from the informal is very significant in providing adequate explanation facilities in knowledge-based systems. In particular, a continuing requirement for a high level of knowledge base maintenance activity is now seen as characteristic of knowledge-based systems, and the linkage of computational representations to their underlying sources is seen as critical to effective maintenance. Thus, the emphasis of current research on knowledge acquisition tools has become one of *integration*—to link the informal original sources, their processing, the derived formal structures, and their ongoing application, into a unified system (Shaw & Gaines, 1987; Boose & Bradshaw, 1987; Gaines, 1988; Gaines & Linster, 1990).

Figure 1 shows a partitioning of the 'knowledge base' in such integrated systems which brings out the diversity of data sources and types relevant to standards issues. Note that, while there is a flow from the informal to the formal through various stages of processing, the final knowledge base is shown as retaining the data at all stages, indexing and linking it appropriately. This is important to the provision of effective explanation, to the support of knowledge base maintenance, and to the ongoing knowledge acquisition process.

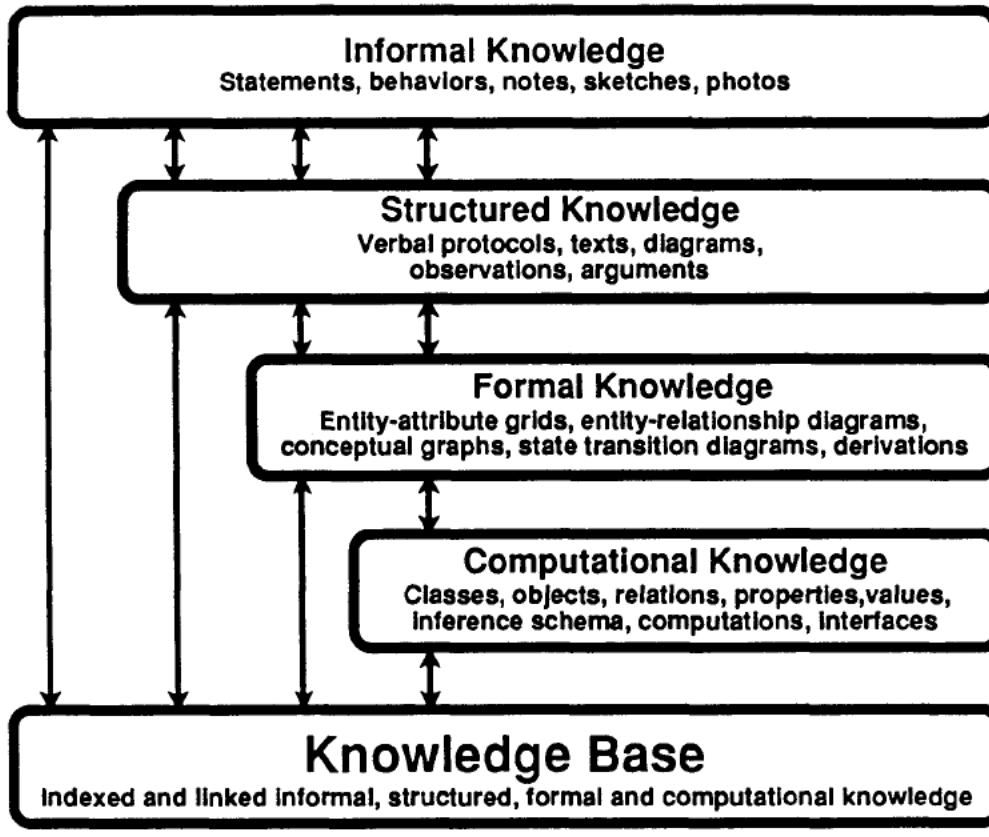

Fig.1 Levels of representation in a knowledge base

RELEVANT STANDARDS

From Figure 1 it can be seen that the data interchange standards involved in knowledge acquisition range from video and sound media, through text, hypertext, specialist mediating representations, to various forms of knowledge bases. There is also a need for inter-application integration where some degree of standardization in communication protocols, application interfaces, script languages, and so on, would be valuable.

The majority of components are ones where the standards are already established in other disciplines, such as those for images and text, or are being established now, such as those for hypertext and knowledge bases. Even in the few components that are specialist to knowledge acquisition and its mediating representations, it may be that knowledge acquisition standards are not appropriate, that one should be looking more for well-defined data structures within other standards, e.g. repertory grids can be directly represented in various knowledge bases. In all cases it would be advantageous to have well-defined preferred use of existing standards for different knowledge acquisition functions.

The following subsections are a preliminary list of relevant requirements and standards.

Video and Sound

The standards for magnetic media, CD's and videodiscs are well-established. Timing information in SMPTE format is important on tapes to allow them to be synchronized with one another and other event recording equipment.

There is increasing use of digitized audio under direct computer control without well-established standards. One can expect similar use of digitized video as JPEG chips and large discs become available.

Text

Text separates into four components:

- Raw text
- Functional markup, such as paragraphs and headings
- Typography
- Layout

The last two are primarily significant for presentation rather than acquisition. Functional markup is important to acquisition since it usually punctuates a document conceptually. The growing mover towards SGML, and the availability of SGML word processors and tools is significant. Most text commences in word processors with proprietary formats. The capability to export and import in standard RTF format is common, and this is the most useful interchange format. However, structure markup may be implicit in RTF, and, even when 'paragraph styles' are used, it may be peculiar to a document.

Graphics

Graphics separates into four types which may be mixed:

- Line art
- Continuous tone
- Half tone

The last two are primarily significant for presentation rather than acquisition. Automatic picture analysis is as yet uncommon in knowledge acquisition systems, although human commentary on presented pictures is important. There is a bewildering variety of 'standards' for pictures making interchange more difficult than it need be.

Line art is a common syntactic level for graphics in many different specialist communities in computing, and there are standards for this level such as GKS and Postscript. However, semantic considerations have led to a variety of standards specialist to disciplines such as IGES and EDIF. These are usually not a problem since the relevant knowledge acquisition tools are domain specific.

There is growing use of line art to carry knowledge in the form of concept maps, formal semantic nets, and CASE formats such as $IDEF_0$. Even though the interchange format may be purely graphic, the interpretation is highly specific and best treated separately as a 'visual language.'

Visual Languages

Graphic presentation, entry and editing of knowledge structures is common in knowledge acquisition and becoming an increasingly important component of many tools. There are four major types of visual language currently in use:

- Specialist data entry and editing
- Concept maps
- Semantic networks
- CASE formats

The first type is specific to a tool and standardization is not important. The last has a range of well-defined formats which are significant for the final stages of knowledge acquisition but usually too stylized and complex for direct use by experts in the early stages.

Concept maps are an informal representation of concepts and their relations developed in education but now being used on large-scale knowledge acquisition projects. Their syntax is

well-defined but their semantics is deliberately open-ended. Their interchange is becoming increasingly important and standardization would be valuable.

Semantic networks with well-defined syntax and semantics are important as a humanly readable, and editable, presentation of all or part of a knowledge base. Again, their interchange is becoming increasingly important and standardization would be valuable.

Hypertext and Hypermedia

Hypertext and hypermedia are widely used for data capture and storage during the early stages of knowledge acquisition. Currently there are no interchange standards, although NIST has a working group on hypertext standards, and interchange requires common use of proprietary tools.

Hypermedia structures are heterogeneous and open-ended, and the technology is new and rapidly developing, so that it is difficult to imagine useful standards in the near future. It would be useful if knowledge acquisition tool developers separated their data structures from those of hypermedia shell so as to make interchange possible at a semantic level.

Databases

Databases are a common source of material in knowledge acquisition, and passing around large datasets originating in databases is common, as it is in the machine learning community. There are a number of common standards which have been in use for some years, and reading them, and converting between them, is not a significant problem.

Knowledge Bases

Most knowledge acquisition tools have an export module as their final component that generates knowledge structures in the format required for some expert system shell. Here there are no standards and it is necessary to have a different module for every shell supported. To some extent this can be treated as a syntactic process to be carried out through scripts, but there are also major semantic differences in the functionality of shells that makes the application of a knowledge structure to problem-solving a different task for every shell.

Thus the export module is not just a software engineering task of data conversion. It is rather a knowledge engineering task of attempting to guarantee that, independent of the domain and type of application, the tool having acquired the appropriate knowledge will deliver an effective problem-solving system.

At the current state of development of knowledge-based systems these problems may be expected to remain for many years. However, standards initiatives such as KIF are important in providing at least one widely used, non-proprietary knowledge representation format to which knowledge acquisition tools should export.

Since knowledge acquisition tool developers are already in a situation of exporting to many different performance tools they are in a good position to make a critical evaluation of KIF proposals relative to what is already done, and what seems to be the minimal essential requirements.

It would not be reasonable for first-generation KIF standards to support the full range of requirements expressed in Figure 1—they are targeted on the 'computational knowledge' component. However, it would be valuable in linking this component to the others if any KIF made provision for structured annotation. This could be discarded by basic performance engines, but the same knowledge structures could then also be used in more complex knowledge support systems, for example, in providing hypermedia explanation.

An open-ended syntax providing for annotation is a simply implemented requirement. The SGML standard provides a good model in this respect.

Inter-Application Communication

Inter-application communication is probably not perceived in the same way as data interchange. However, knowledge acquisition tool benches are increasingly being developed through the heterogeneous integration of separately developed tools. In particular, this applies to the integration of both knowledge acquisition and performance tools, and to knowledge acquisition and hypermedia tools.

Our current generation of interactive tools of all types has become so dependent on visual interaction through mouse and graphic interfaces that the need for a parallel script language for inter-application integration is often forgotten. It would be useful to develop a culture in which it is accepted that all actions performable by an interactive user can also be performed by another program through an inter-application protocol.

CONCLUSIONS

Research and development activities in knowledge acquisition are now at a stage where lack of standards is a definite impediment to progress. There is a high volume of activity worldwide and many significant technologies have been developed. However, very few institutions are able to support the full range of developments required, and none are in a position to undertake projects in a comprehensive variety of domains. The interchange of datasets and the sharing of tools is vital to further progress. The effective use of existing standards and awareness of, and involvement in, the further development of relevant standards is important to the knowledge acquisition community.

ACKNOWLEDGEMENTS

The work at Calgary was funded in part by the Natural Sciences and Engineering Research Council of Canada. We are grateful to many colleagues who have played significant roles in exchanging ideas and research data with us, notably, Jeff Bradshaw, Bill Clancey, Paul Compton, Bob Jansen, Marc Linster, Ross Quinlan, Alain Rappaport, Mildred Shaw and Brian Woodward, plus many stimulating and influential discussions at the Knowledge Acquisition Workshops worldwide.

REFERENCES

Boose, J.H. & Bradshaw, J.M. (1987). Expertise transfer and complex problems: Using AQUINAS as a knowledge acquisition workbench for knowledge-based systems. **International Journal of Man-Machine Studies, 26,** 328.

Boose, J.H. & Gaines, B.R., Eds. (1988). **Knowledge Acquisition Tools for Expert Systems.** London, Academic Press.

Boose, J.H. & Gaines, B.R., Eds. (1990). **The Foundations of Knowledge Acquisition.** London, Academic Press.

Gaines, B.R. (1988). Knowledge acquisition systems for rapid prototyping of expert systems. **INFOR, 26**(4), 256-285 (November).

Gaines, B.R. & Linster, M. (1990). Development of second generation knowledge acquisition systems. Wielinga, B., Boose, J.H. & Gaines, B.R., Schreiber, G., van Someren, M., Eds. **Current Trends in Knowledge Acquisition.** pp. 143-160. Amsterdam, IOS.

Shaw, M.L.G. & Gaines, B.R. (1987). KITTEN: Knowledge Initiation & Transfer Tools for Experts & Novices. **International Journal of Man-Machine Studies, 27,** 251-280.