[MT Summit IV, July 20-22, 1993, Kobe, Japan]

From Electronic Dictionaries to Very Large-Scale Knowledge Bases

Toshio Yokoi
Japan Electronic Dictionary Research Institute, Ltd. (EDR)
Mita-Kokusai Building Annex, 4-28, Mita 1-chome, Minato-ku Tokyo 108, Japan
Tel: +81-3-3798-5521, Fax: +81-3-3798-5335
E-mail: yokoi@edr.co.jp

1. Introduction

With the computer reaching a period of maturation of sorts, a movement toward software technology as the next fertile ground for growth and development has become prominent. Apart from individual software applications, this also involves technology, relating to information and to knowledge, as the foundation on which advanced software technology of the future must be built comprehensively. Technology for very large-scale knowledge and very large-scale knowledge bases, and for the construction of such knowledge bases will be the concrete manifestation relating this technology to knowledge itself.

If very large-scale knowledge bases are perceived from the narrow viewpoints of their roles in facilitating the development of expert systems or of enlarging the scale of such systems, then the significance of such knowledge bases will be diminished. It is necessary to determine the proper place of such knowledge bases. This can be achieved through a close examination of the future development of information processing technology, which has reached a turning point, and through consideration of the most reliable approach to artificial intelligence, acknowledging that the expectations engendered by the "AI boom" have been somewhat inflated. This paper attempts to describe very large-scale knowledge bases built on an amalgamation of language processing and knowledge processing.

2. The Concept of the Very Large-Scale Knowledge Base to Be Sought

In this section the basic aspects of very large-scale knowledge bases are discussed. These include the characteristics which the desired very large-scale knowledge base must have, and what role it will play in information processing and artificial intelligence.

Among the conditions which must be met, emphasis is placed on the role of natural language. In the field of artificial intelligence, much research on natural language has in the past dealt with the human interface or with interface media. In this paper, however, natural language is considered to be the medium to represent knowledge or to be the knowledge representation language. This difference is crucial. In order to clarify basic principles, this paper will explain the significance of entrusting natural language with a large part of the role of knowledge representation and will also explain its relation to other media; because the way in which separate media are used to represent knowledge will determine the fundamental nature of very large-scale knowledge base technology.

There are four points to be made, as follows:

(1) Integration of Different Approaches

The approaches from different fields each have their advantages and disadvantages. The fact that similar attempts have been made from a variety of fields demonstrates that the time is ripe to treat very large-scale knowledge bases as a major theme of research. However, another important question is whether very large-scale knowledge bases can be set as a research theme which combines the advantages of the various approaches.

From the standpoint of clarification of all the knowledge possessed by human beings, for the present only the most preliminary stage of research can be undertaken. By expanding the scale of the system, the resulting knowledge base will of itself acquire a considerable inertia. Hence the very large-scale knowledge base to be developed must support and accelerate the further elucidation of knowledge and the continued growth of knowledge bases; it must under no circumstances impede these processes. The fact that a variety of approaches can be integrated serves to ensure this. In other words, the most important conditions to be met are the ability to expand the scale of the knowledge base without serious difficulties, the absence of any serious problems resulting from an increase in scale, and most importantly, enhancement of the value of the knowledge base through an expansion in scale.

Considered in this way, the desired large-scale knowledge base will adopt an approach of knowledge processing in order to achieve the certainty of a framework for knowledge management, and an approach emphasizing multimedia for the broad range of the knowledge to be handled; and these are connected by an approach based on document processing and natural language processing.

(2) Assignment of an Appropriate Role for Knowledge Representation Media

The medium that is used to represent knowledge determines the basic character of that knowledge. In order to appropriately integrate the various approaches to very large-scale knowledge bases, roles for the knowledge representation media to be used in each approach must be assigned appropriately from the standpoint of integration. For this reason, as has been common in AI, it is insufficient in terms of knowledge representation to refer only to so-called knowledge representation languages or knowledge representation languages in the narrow sense of those which are used in expert system shells. It is necessary to consider in an integrated way all the media that human beings have contrived in their long history. This is because even if knowledge cannot be understood by a computer, its worth will still be evident so long as it is understood by a human being using a computer. Of course, computers must be made to understand enough to enable them to support the understanding of human beings. In other words, knowledge should be represented to facilitate interpretation by both humans and computers.

Very large-scale knowledge bases must take all of the following knowledge representation media into account.

Natural languages:	Japanese, foreign languages
Formal languages:	Algebraic formulas, logical formulas, programming languages, knowledge representation languages (in the narrow sense), etc.

Picture languages: Architectural design drawings, electronic circuit diagrams, music scores, etc.
Images: Static images, dynamic images, animation, etc.
Sounds: Speech, music, and any other type of aural information

Each of these media has its own role to play, and boasts its own unique expressiveness, which cannot be replaced by another. And, by combining them appropriately still greater capabilities can be obtained.

The roles of each of these media will be considered as appropriate, but an important role will be assigned to natural language in particular. This is because natural language is the medium with the most versatile expressive capacity and with the most highly developed ability for symbolic representation. This medium, with the most general yet powerful expressive capability, we position at the center of very large-scale knowledge bases.

(3) Emergence as a New System Architecture

A very large-scale knowledge base must not be relegated to the status of a subsystem which handles the knowledge-base components of the computer system, or which only serves to accumulate and manage knowledge. A very large-scale knowledge base or a very large-scale knowledge base system must constitute a generalized application system or generalized expert system. Put more strongly, the very large-scale knowledge base must amount to a new proposal for computer system architecture.

The architecture of a computer system is in essence stipulated by the language which is at the core of the representation of information and knowledge, that is, by the programming language used. Here programming languages present all the computer media. By raising programming languages to a more sophisticated level, the system architecture likewise becomes more highly developed. Representative of this are the debates over the past decade or so regarding logic programming, functional programming, and object-oriented programming.

The adoption of information and knowledge representation media in very large-scale knowledge bases centered on natural language amounts to a proposal for the adoption of natural language to serve as the main medium of computer systems. This means making natural language the central language of the computer as well. In response to such a proposal, generally the following kind of question is raised. "Does this mean that from now on natural language will be used to write programs?" No, it does not. It means, however, that our conceptions must be altered. The aim is not to raise the language levels from the computer side so as to approach or attain natural language. Rather, the computer is approached from the side of the information itself which the computer handles. In other words, we have natural language on the information side, and programming languages on the computer side. From the knowledge side, the system architecture is specified by natural language; whereas from the computer side, the system architecture is specified by programming language.

(4) Broad, Shallow, Robust Processing of Natural Language

Languages available for the representation of knowledge include natural languages and programming languages. Knowledge represented using programming languages can be accu-

rately executed, processed and understood by a computer without modification. Knowledge represented through natural languages cannot similarly be operated on by a computer. This much is clear when one considers the tardy progress made by the theme of natural language understanding in artificial intelligence.

A comprehensive understanding of all phenomena relating to natural language is essentially tantamount to a comprehensive understanding of human intelligence. The time required to bring to completion our research on the theme of natural language understanding must needs be measured in centuries. Hence the functions for natural language understanding and for natural language processing which we are here undertaking, or can undertake for the foreseeable future, must far precede any such end of inquiry. Under such circumstances, is it possible to place such emphasis on natural language?

At present, technology for natural language processing is, up to a certain level, on its way to becoming technology which can be used reliably. Processing technologies in the area ranging from morphological processing to syntactic processing have nearly reached the threshold of practical applicability. In the areas ranging from semantic processing to context processing, somewhat more effort is needed. In the fields from semantics to context, we must not place expectations on "deep" processing such as is discussed in natural language understanding. The processing must be, above all, shallow, yet must be capable of application over a broad range.

Through the above four points, the conditions to be targeted for very large-scale knowledge bases have been summarized; the relation between knowledge representation media and knowledge in such very large-scale knowledge bases is summarized in Fig. 1. It must be pointed out that in the figure, the "programs for humans" correspond to the "multimedia data" part of knowledge for computers, and the "knowledge for computers" corresponds to the "source texts" part of the programs for humans.

3. EDR Electronic Dictionary

3.1 Overall Structure of the Dictionary and the Roles of Each Section of the Dictionary:

The EDR Electronic Dictionary has been developed for Japanese and, as a representative of foreign languages, for English. Their overall configuration and the relations between component dictionaries are shown in Fig. 2. They can be broadly divided into three layers. The Concept Dictionary holds "deep-layer" information; its purpose is collect knowledge and bring the computer to understand the meanings of words. The Word Dictionary, Bilingual Dictionary and Cooccurrence Dictionary provide "surface-layer" information, and are meant to teach the computer syntactic and morphemic behavior. The Word Dictionary also has the role of linking the "surface" and "deep" layers. The EDR Corpus and the Text Base are collections of sample text data meant to serve as materials for the development of the dictionary. Each section of the dictionary is related to each other. The solid lines indicate how each section of the EDR Dictionary shares data items with other sections; the dashed lines indicate how some sections of the Dictionary are developed by using the data. The Fig. 3 illustrates the mutual relations between the Concept Dictionary, Word Dictionary and the Text Base.

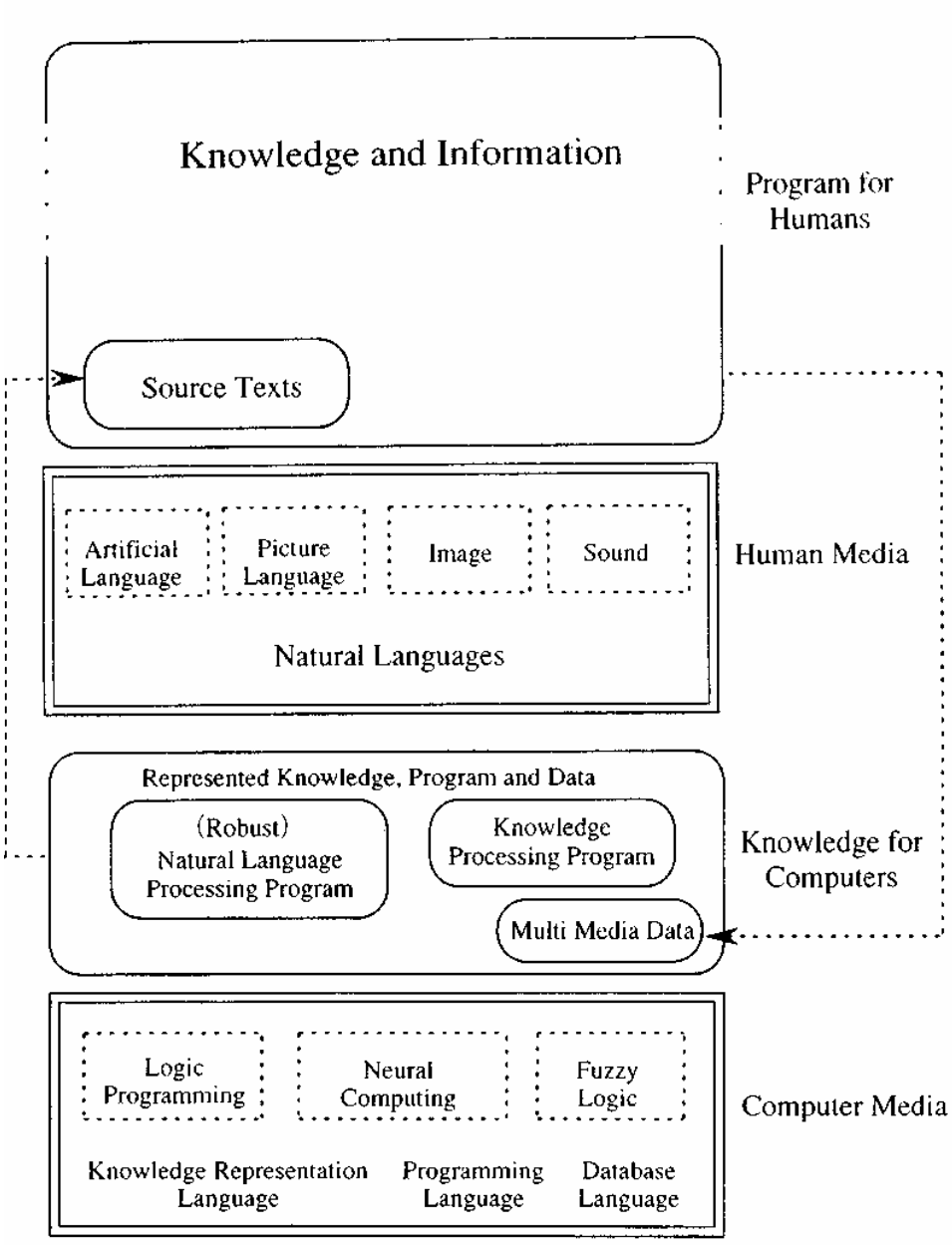

Figure 1. Representation Media and Knowledge

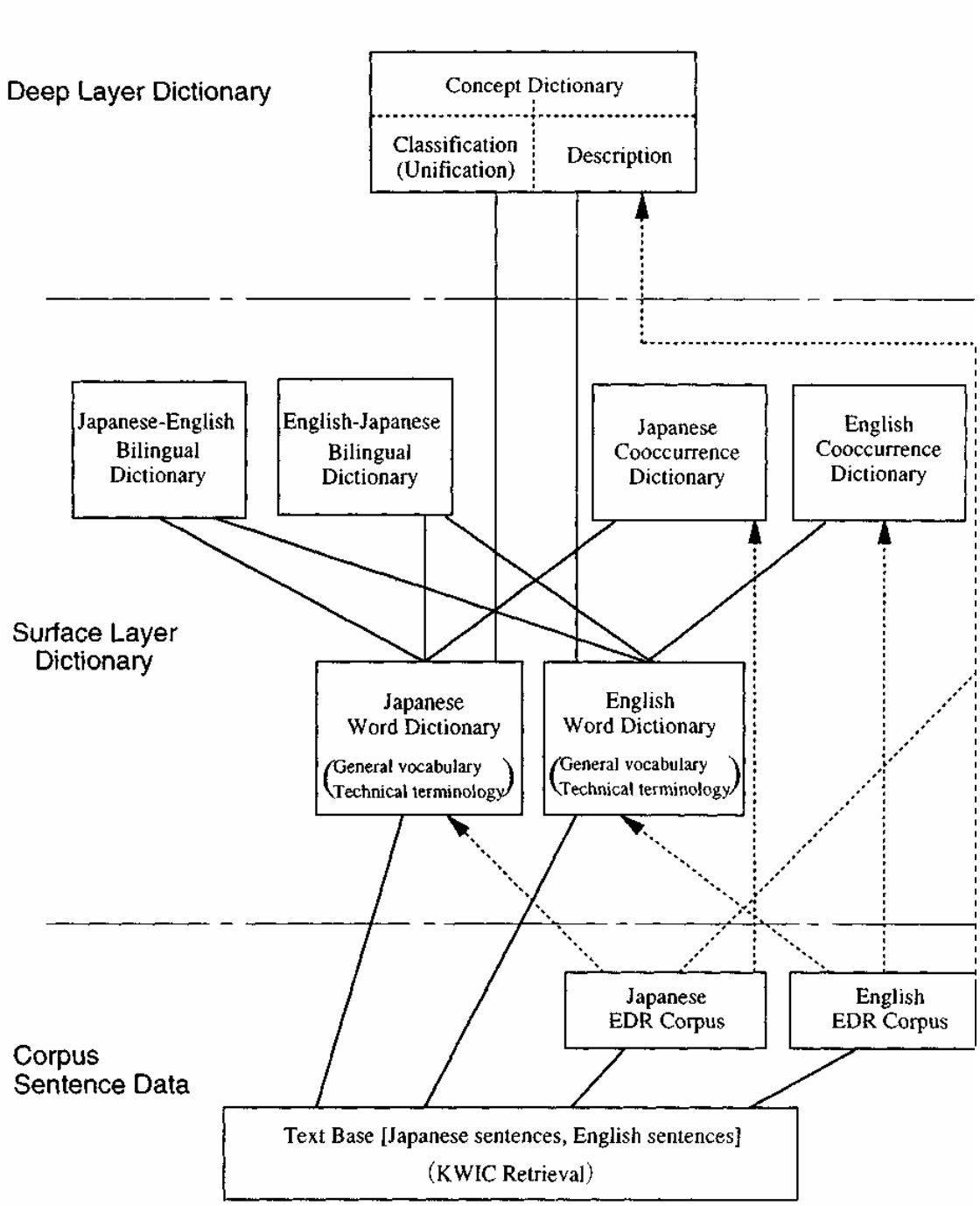

Figure 2. EDR Electronic Dictionary

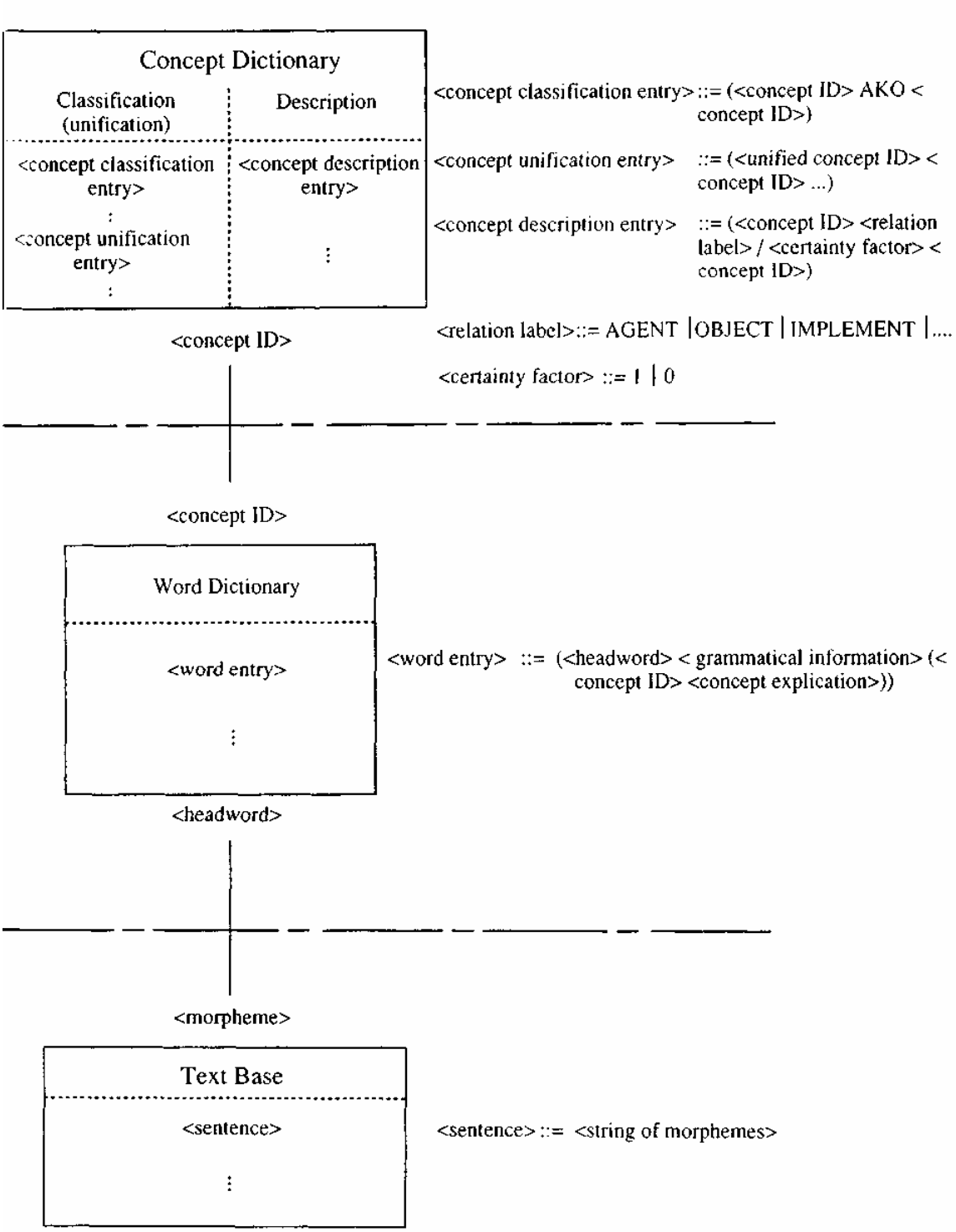

Figure 3. Internal Structure of the EDR Dictionary

3.2 The Electronic Dictionary as a Very Large-Scale Knowledge Base

The nature and the role of the EDR Electronic Dictionary can be explained in terms of the desired properties of a very large-scale knowledge base (Section 2 above). This is a matter of allotting an appropriate role to the knowledge representation media, and of establishing a continuous and balanced connection between what can only be understood by humans, and what can be understood by computers. In the explanation reference is made to Fig. 3. Note that there are two types of knowledge representation media used in the electronic dictionary: these are natural language, and a simple formal language, which may be called a dictionary description language, for use in representing the contents of the Concept Dictionary and grammatical information.

What the meanings of words and phrases are, and how words should be defined, are matters which are addressed by a great number of theories in the field of linguistics. One such representative theory holds that ultimately the most precise way to define a word is to display sentences containing that word. In keeping with this reasoning, the EDR Electronic Dictionary accumulates as great a number of examples as possible in the Text Base, and can search for sentences containing each word in the Keyword in Context (KWIC) format. The Text Base provides the most accurate definitions of words. However, clearly this means alone is insufficient to produce understanding on the part of the computer. Such definitions, which can only be understood by human beings, will serve as an important foundation for the gradual improvement of the precision of the Electronic Dictionary as a whole.

As part of preparations to enable a computer to understand the meanings of words defined by the Text Base, in the Japanese and English Word Dictionaries the meanings are divided into what are called concepts. A multiple number of concepts, corresponding to the polysemity (multiplicity of meanings) of a word, are assigned to the word. Concepts are defined by concept IDs and concept explications. Concept IDs are prepared for computers to identify concepts. Concept explications are for the benefit of human beings, and are expressed in natural language sentences so as to specify the concept as unambiguously as possible. The meaning of a word is distributed continuously, and fluctuates each time it is used. The treatment of meanings as a certain number of discrete concepts is an approximation in order to enable computers to understand them.

Next, concept explications must be expressed so that they can be understood by computers. To this end, the Concept Dictionary adopts the method of listing the correlations between concepts. The way the concepts are related in concept classification and concept unification will enable computers to understand whether two concepts are equivalent, and if not, how close they are to each other. Sentence concepts (called concept-relation representations) are then defined as extensions of word concepts. The purpose of concept descriptions is to enable the computer to judge the reasonableness of the concept of a sentence, to convert it into the concept of a similar sentence, and to judge the degree of similarity between sentence concepts. But because the relations between concepts are limited to approximately 40 different kinds of relations, and because the way in which these relations should exist is limited by the precision of the contents of the Concept Dictionary, here again a considerable amount of approximation is employed.

Between the Text Base and the Concept Dictionary, meanings are subjected to several stages of approximation. Among the meanings originally associated with a word, the computer is able to understand only a small fraction. But in light of the precision of natural language processing in the

foreseeable future, even this much is sufficient. However, through efforts to raise the precision of approximations and new innovations in methods of approximation, the precision of the computer's understanding should be gradually improved. The Concept Dictionary is thus a collection of the simplest sentences comprehensible to the computer, a kind of text base for the computer, as it were.

3.3 The Dictionary Development Process as Lexical Knowledge Acquisition

Additional major goals of the EDR Electronic Dictionary Project are the establishment of a methodology for the development of electronic dictionaries and the implementation of a development support system. The development of a dictionary strongly suggests the notion of great human labor. Of late, as in so many other fields, computers have come to be used, but only in data management and printing processes. Of course, the development of the EDR Electronic Dictionary also required considerable human effort. In order to refine the dictionary sufficiently, so that it may serve as a foundation for the development process, use had to be made of human labor. At this stage the main point of concern is the arrangement and management of the overall process so as to obtain a uniform and high-quality result. Thus a series of tasks were undertaken, including the creation of worksheets suited to external workers using a computer, design of work processes, validation of dictionary data, and the subsequent dictionary development by researchers.

But if the dictionary development reaches a certain point, then techniques and systems somewhat more suited to large-scale lexical knowledge acquisition can be created. At present, a system with the configuration shown in Fig. 4 is being prepared. For the time being, the system will be called the Integrated EDR Electronic Dictionary Development System. The programs and systems developed and used to date will be organized into what may be properly called an integrated lexical knowledge acquisition system, albeit at an early stage.

The dictionary data extraction function analyzes large amounts of text data (the Text Base), extracts unregistered words for the Word Dictionary, infers grammatical characteristics, and generates analytical data on morphemes, syntax and semantics for the EDR Corpus. It is a highly functional natural language processing program, which is executed on the EDR Electronic Dictionary. Hence the dictionary precision is improved, and as the scale is expanded the precision of the extracted data is also enhanced, while at the same time the amount of human intervention required is gradually decreased.

The editing support function supports editing and browsing of the dictionary at a sophisticated level, and this function is also advancing toward more intelligent capabilities, including automatic checking of constraints on dictionary data items. Functions are nearly the same for editor and for user. The user functions are somewhat less extensive compared with the editor functions, but to facilitate creation of a variety of application dictionaries a wide variety of data formats can be handled. The figure addresses functions, and so is drawn as if editor and user access the same dictionary, but of course system configurations are entirely different.

The function of the communication of dictionary information mediates the mutual exchange through a network of information relating to dictionaries between editor and others (users and external workers).

Preparations are underway with the aim of enabling this system to be used long after project

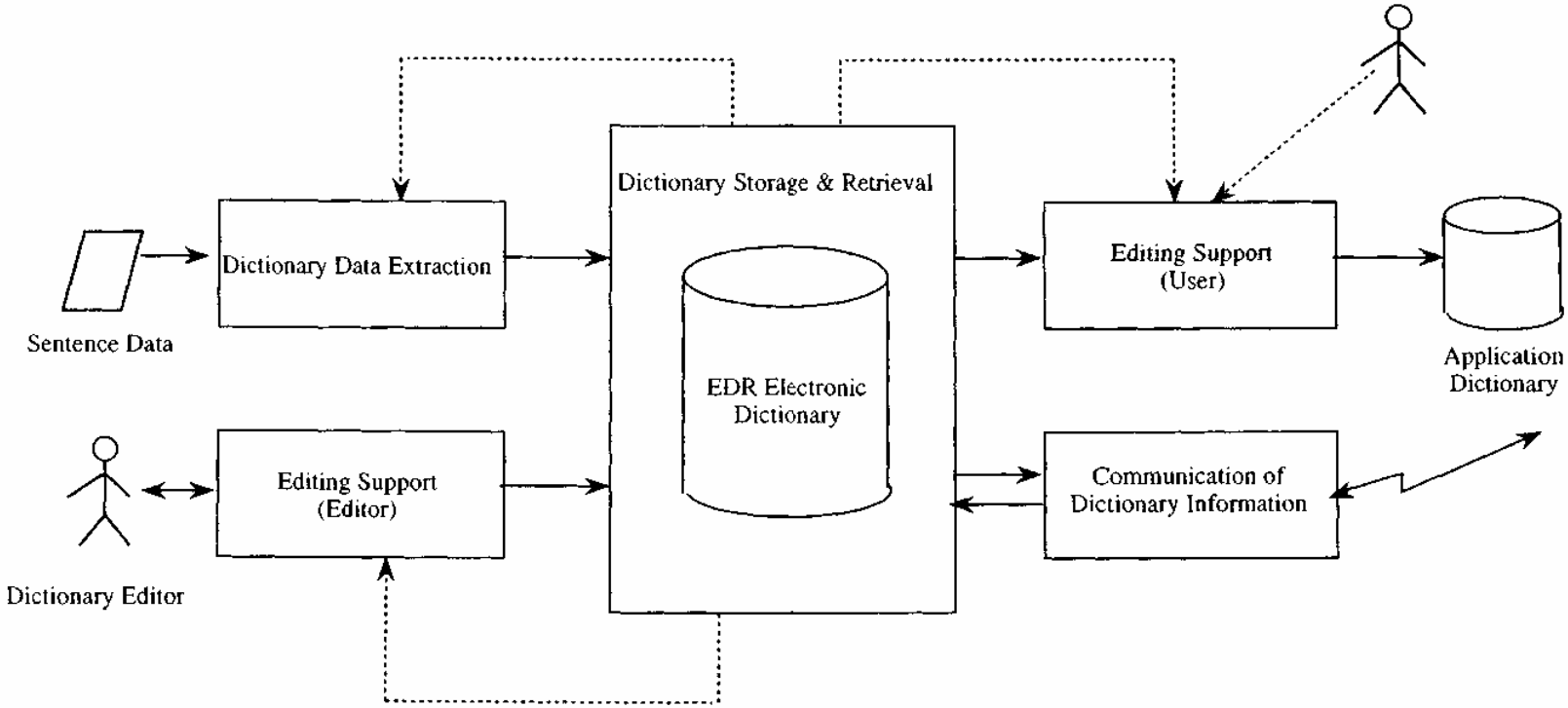

Figure 4. Functions of the Integrated EDR Electronic Dictionary Development System

completion. In future, it will no doubt develop into what may be called a dictionary generation plant, a dictionary factory, so to speak.

4. Plan for Research and Development of Very Large-Scale Knowledge Base (Knowledge Archives)

Efforts are underway to develop a project which would give concrete form to a very large-scale knowledge base as described in Section 2 above in what are called "Knowledge Archives". This paper concludes by introducing a part of the plan.

The Knowledge Archives represent an attempt to come to grips with technology for knowledge itself, that is with new integrating software technology. At the center of this is knowledge processing and natural language processing. In the past, artificial intelligence research has addressed the two either through natural language interfaces for knowledge-based systems, or by using knowledge bases in natural language understanding. By viewing the two as equal and integrating them an attempt is made to develop new technology. It would be bold to call this natural language processing-based knowledge processing.

4.1 Knowledge as Raw Material

In the knowledge archive, only that knowledge which can be explicitly represented and objectively observed is included. This knowledge is information which has been represented in an organized way in a knowledge representation medium.

The knowledge representation media comprehend all media for information representation that can be handled by a computer, including natural languages, formal languages, graphical languages, images, and sound. Each of these knowledge representation media has a role to play, and possesses its own unique capabilities, which cannot be supplanted by another. Hence they are treated equally in such a way that each of their roles can be fulfilled. However, from the standpoint of knowledge processing, two of these are chosen to be kernel media based on the criteria of generality of expressive capabilities and sophistication of symbolizing capabilities. These are natural language (Japanese), which meets these criteria from the human side, and knowledge representation language (a language corresponding to the knowledge base functions of the Knowledge Archive), from the computer side.

Knowledge documents represent knowledge in an appropriate knowledge representation medium, and enable its objective observation and analysis. Knowledge documents are an expression of knowledge directed to the composite entity of human beings and computers. It was mentioned that all knowledge representation media are to be treated equally, but an appropriate prioritization is needed for the purposes of actual research and development. In order to take this into account the following three steps are being considered.

<1> Knowledge documents in the Japanese language.

Knowledge documents in other knowledge representation media are replaced by knowledge documents in Japanese with approximately the same meaning, to render the entire archive uniform. When using Japanese, in some cases controlled Japanese (Japanese with constraints added to facilitate the extraction of knowledge) is employed.

<2> Knowledge documents in Japanese and formal languages.

Among the formal languages considered are knowledge representation languages, programming languages, algebraic equations and logical expressions. Chief among the knowledge representation languages is that of the system itself, and representative languages are included insofar as possible. Functions for processing and understanding documents which are a mixture of both Japanese and formal languages represent a new theme for research.

<3> Knowledge documents in Japanese, foreign languages, formal languages, picture languages, images and sound.

Attempts will not be made to realize extremely general functions for processing and understanding knowledge represented in graphical languages, images or sound. Rather, the goal is a simple framework enabling utilization of the characteristics of the knowledge documents in question.

In order to establish technology for knowledge archives, knowledge fields representing a variety of the features of diverse types of knowledge must be chosen, and research and development conducted on a considerable volume of knowledge documents. To this end, mutual cooperation with the various organizations which hold such knowledge documents and which possess a wealth of experience is vital. With respect to knowledge documents in different foreign languages, this implies mutual cooperation with the research groups of the countries in which the languages are used. The following are the types of knowledge documents being considered.

(1) Text

This consists mainly of natural languages (Japanese and some English), but graphics, images (photos, etc.) and other knowledge representation media are also used. The documents themselves include scientific and engineering papers, newspaper articles, patent documents, legal documents and judicial precedents, and manuals.

(2) Knowledge representation (through knowledge representation languages) and programs

These rely mainly on formal languages, namely knowledge representation languages and programming languages. Specifications are also included; the knowledge representation media used in the specifications include Japanese, formal languages (specification description languages, algebraic equations, logical formulae, etc.), figures and tables.

(3) Data

This employs formal languages in the form of database languages, as well as Japanese and other media for specifications. Extraction of knowledge from massive data constitutes a major theme to be researched.

4.2 Structure of Knowledge

In this paper, only the fundamental aspects of knowledge structure are introduced, including the manner in which the contents of knowledge documents are accumulated in a knowledge base and

put to use, and the type of structure of the knowledge. In the explanation which follows we use an analogy with the EDR Electronic Dictionary (Section 3) to aid an intuitive understanding. It is not the case that knowledge archives are constructed in this manner in actual practice.

The EDR Electronic Dictionary holds knowledge relating to words (lexicon) in natural language. The Knowledge Archives holds sentences in natural language and in other knowledge representation media, as well as knowledge relating to texts and paragraphs. This is the essence of the analogy. In abridged form, the structure of knowledge is illustrated in Fig. 5, following the analogy with the EDR Electronic Dictionary in Fig. 3. We omit an explanation here, but expect that the reader can gain a rough understanding of its meaning.

4.3 Functional Configuration

The functional configuration of the Knowledge Archives is summarized in Fig. 6. This figure also explains the archive functions, but does not indicate the structure of the system itself. The Knowledge Archives function as the most universal expert system.

The knowledge extraction functions serve to extract high-quality knowledge from large volumes of knowledge documents, automates their accumulation in the knowledge bases insofar as possible, and provides efficient support for the archives. Here extraction of knowledge from knowledge documents involves the creation of knowledge documents which are condensed descriptions of the original documents. If unlimited use of the knowledge representation media were permitted, efficient automated extraction of high-quality knowledge would be difficult. Hence specifications for knowledge representation media with appropriate limits applied, which enhance the "extractability" of knowledge, were devised.

The support function for creating knowledge documents helps experts create new knowledge documents efficiently, and also helps users efficiently create application-oriented knowledge bases of their own.

The knowledge storage and retrieval functions consist of knowledge base functions with self-organizational capabilities for the systematic accumulation of large amounts of knowledge; basic knowledge common to all fields; and shared knowledge which is shared among fields.

The knowledge translation and communication functions support the translation of large volumes of knowledge documents, and enable their widespread and efficient transmission and exchange.

5. Conclusion

While somewhat out of sequence, the need for very large scale, and the importance of very large scale are explained as the conclusion. It is summarized in the following four points:

(1) Although the knowledge concerned is simple, a computer processes large amounts of knowledge at high speeds and with high reliability; human beings use such computers as tools to engage in higher-level intellectual activities. This is a natural division of roles, and will remain unchanged for considerable time into the future. Hence in order to enable computers to offer a high degree of utility, it is not sufficient that they process small amounts of knowledge in a somewhat complicated manner; rather, large amounts of knowledge, vast amounts of informa-

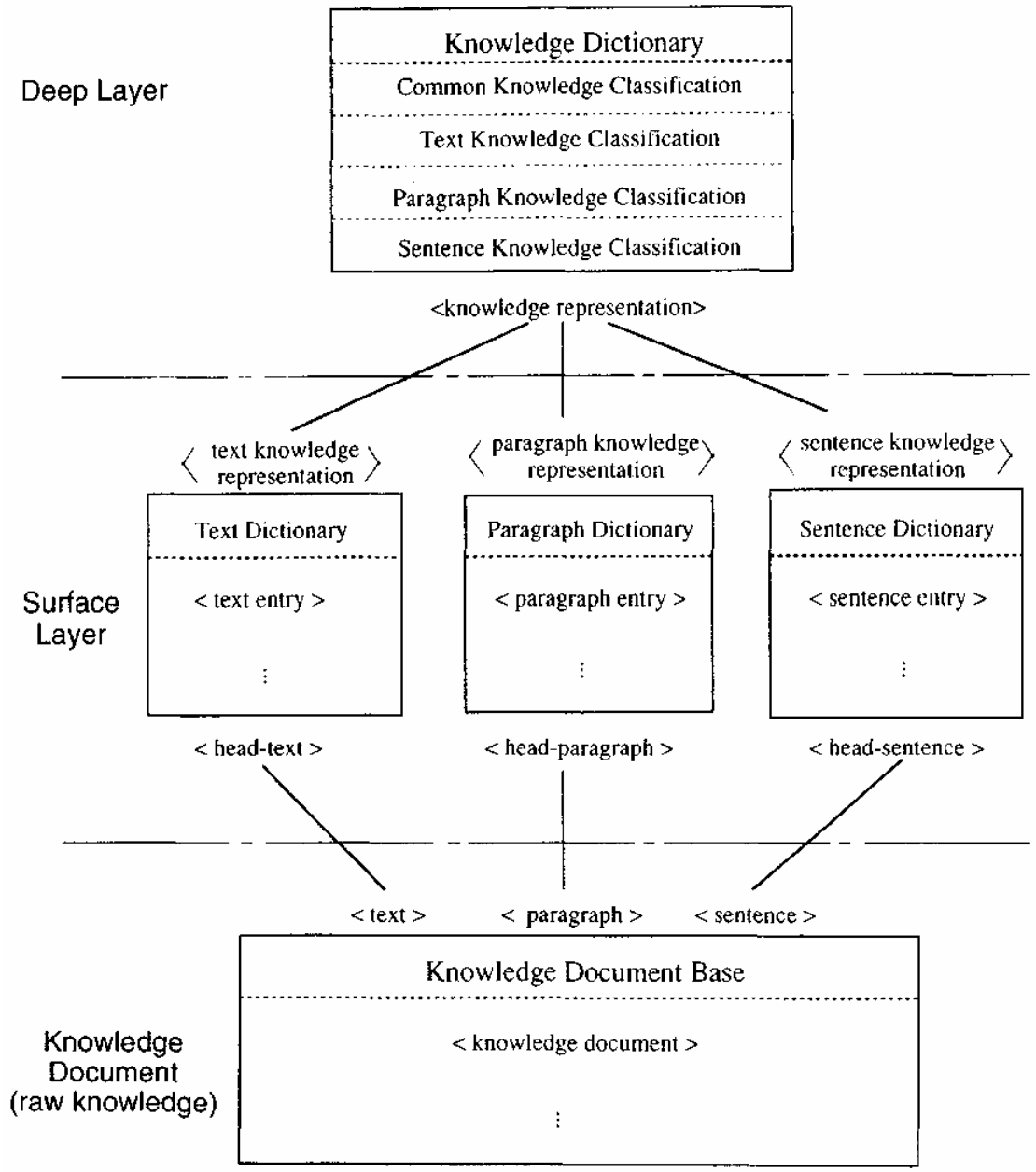

Figure 5. Knowledge Structure of the Knowledge Achives (Analogy to Electronic Dictionary)

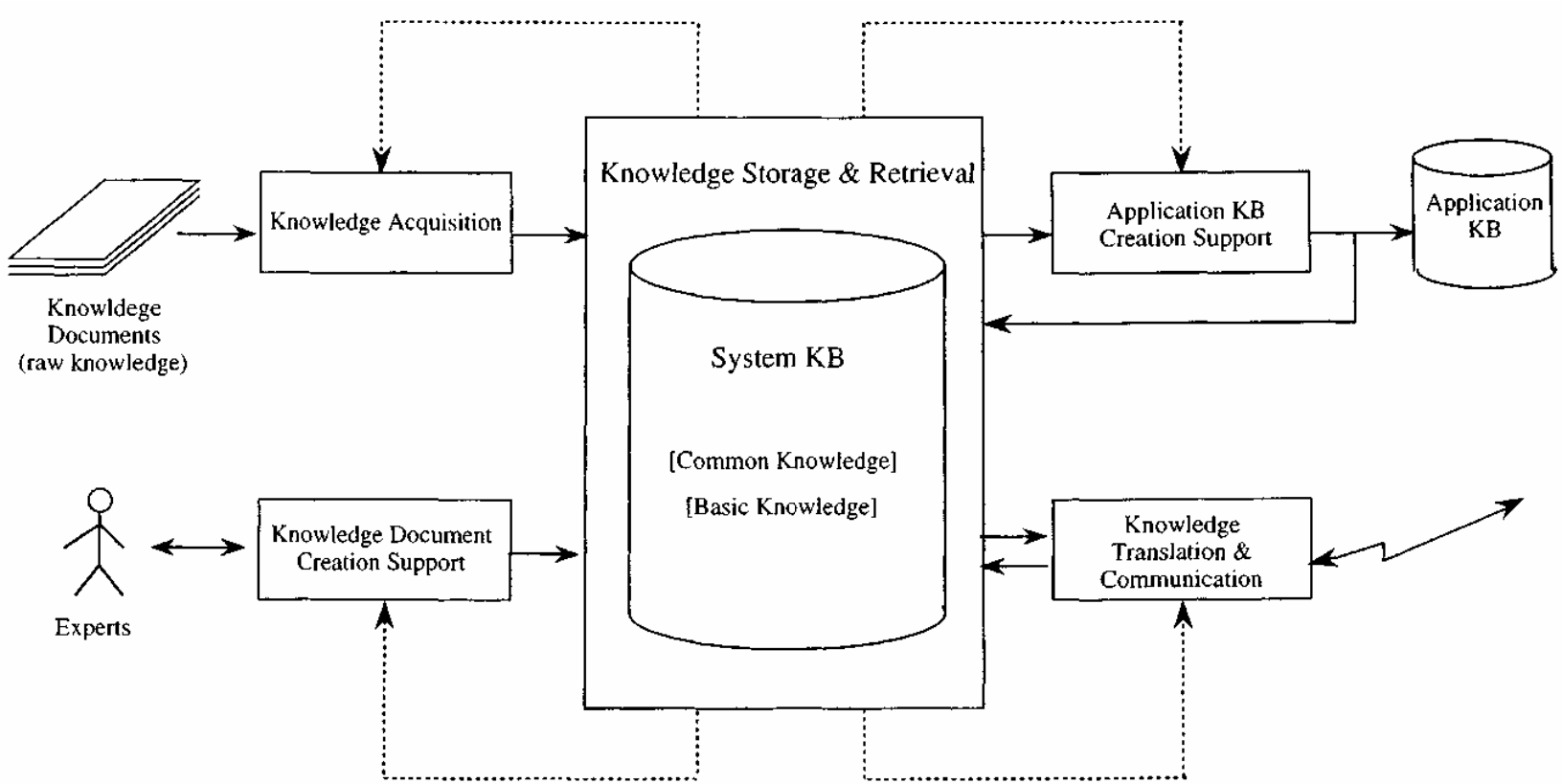

Figure 6. Functions of the Knowledge Archives

tion are indispensable.

(2) This very large-scale technology has two aspects: the technology for dealing with the very large scale of the knowledge itself, and the technology for very large-scale knowledge processing. Efforts to develop this very large-scale technology have in the past been greatly slanted toward very large-scale processing, as born witness by work on massively parallel computers. Very large-scale knowledge is not simply a matter of collecting a large amount of data; it requires an entirely new technology. The time has arrived to grapple in earnest with the very large scale of knowledge itself, both in order to make effective use of what we already have, and also to determine the direction of new processing technology. Among the new technologies for very large-scale processing now attracting attention are memory base inference, which is a parallel knowledge processing method[8], and technology for the application of neurocomputing to massive symbol manipulation. It should be noted clearly that serious R&D on these technologies is possible only by virtue of the existence of large-scale knowledge.

(3) If the raw materials for very large-scale knowledge do not exist on a large scale, then the notion of very large-scale knowledge will remain a dream. Raw materials for the creation of knowledge are databases and text bases. In the past, considerable costs were expected in developing such databases and text bases. However, the spread of word processors and electronic publishing has greatly altered this situation. Large amounts of data are being input into computers every day. Databases have grown large, and already expert guidance is needed in database searches as well. The large-scale accumulation of raw materials for knowledge, as well as the clarification of the limits to simple database retrieval technology, both lead to expectations placed on very large-scale knowledge bases.

(4) While it is spoken of giving the computer only simple knowledge, the ultimate goal is still to have the computer understand complex knowledge and to realize a system with intelligence rivaling that of human beings. To this end, means must be prepared for the steady elucidation of the mechanism of human intelligence. Simple theories and insignificant experiments will not allow access to the heart of diverse knowledge phenomena. Actual knowledge phenomena will have to be grasped and very large-scale experiments will have to be performed, in a protracted effort, to clarify the ecology of knowledge. Very large-scale knowledge, very large-scale knowledge bases and very large-scale knowledge base systems are the experimental tools which will provide us with such opportunities.