Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 59-70. doi: 10.7152/acro.v1i1.12464

Basic Taxonomic Structures and Levels of Abstraction

Marta J Fernandez and Caroline M. Eastman

Department of Computer Science, University of South Carolina, Columbia, SC 29208, USA

1. INTRODUCTION

Taxonomic knowledge structures are often used to organize information. We compare basic taxonomic structures in four areas: thesaurus construction in information retrieval, semantic data models in database management systems, semantic networks in artificial intelligence, and mental structures in cognitive psychology. We then discuss levels of abstraction, in particular the importance of intermediate levels. In mental structures these turn out to be basic levels that are more important cognitively than higher or lower levels. We explore the role of abstraction levels in other taxonomic structures and suggest possible future research in this area.

2. TAXONOMIC KNOWLEDGE STRUCTURES

A taxonomy is one of the most familiar knowledge structures used for organizing information. It consists of a hierarchy of class objects, and it emphasizes the level of inclusiveness of classes (Figure 1). The organization of a taxonomy rests upon the notion of class, where a class embodies a collection of objects sharing certain properties. For example, all birds have wings, feathers, and beaks. Sparrows, canaries, and eagles share these properties and form part of the class of birds. If classes of objects are grouped into more general classes, we then establish an inclusion relation between the (more general) superclasses and their subclasses. For instance, the class of animals includes the classes of birds and mammals.

Taxonomic knowledge structures can represent objects, classes of objects, and descriptions of these objects. Their descriptive power has been influential in many areas of research. Even the classical, also known as analytical, definition of a term is based on the notion of taxonomy [Evens et al. 83]. First, the *genus* (or class) of the term is specified. Then, the *differentiae* (or attributes) of the term are appended so as to distinguish it from all other members in the class. A bird, for example, can be defined as an animal which has wings and feathers.

In this section, taxonomic knowledge structures in four areas of research are discussed: semantic data models in database systems, thesaurus construction in information retrieval, semantic networks in artificial intelligence, and mental structures in cognitive psychology. The terminology used in these areas of research is by no means standard. Not even the same vocabulary is used within a single field of study; it depends on the researcher. However, rather than impose a standard, we will use the terminology of researchers that are cited.

2.1 Semantic Data Models in Database Systems

A database system provides facilities for storage and manipulation of large amounts of formatted data in a computer. Knowledge representation in the database environment falls under the domain of data modeling.

ISSN: 2324-9773

Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 59-70. doi: 10.7152/acro.v1i1.12464

A number of data models have been proposed. Each provides a different set of capabilities to represent data and its semantics. That is, a property that can be directly represented in a particular data model might require a complex representation in other models or might not be represented at all [Brodie & Manola 88].

Hierarchical and network data models represent objects as nodes and organize them into a tree or network structure, respectively. The relational model organizes objects into relations, or sets of n-tuples. These three types of models, known as classical models, have been criticized for their limited capability to represent effectively a real-world system. The basic problem with these models lies in the fact that they are essentially record-oriented structures. Relationships between objects are limited to those that exist between records in the given data structure. As a result, the expressiveness of these models is too constrained, and the semantics of the data cannot be expressed directly.

In sharp contrast to previous record-oriented models, semantic models provide modeling constructs that strongly resemble real-world phenomena [Furtado & Neuhold 86]. Semantic models were originally used only as a tool for database design. An application was first modeled using a semantic model, and then it was transformed into a classical model [Hawryszkiewycz 84]. This last step was necessary for implementation purposes, since most database systems at that time were based on classical models only. Nowadays, there exist a number of database management systems based on semantic data models. In addition, some systems built upon classical models offer front-ends that support semantic modeling [Hull & King 87].

Semantic data models can be used to view real-world applications in terms of objects and relationships among objects. These models achieve data abstraction by providing four major abstraction relationships to model an application: generalization, aggregation, classification, and association [Brodie & Ridjanovic 84]. Through the first three of these relationships, objects can be grouped into classes, where a class of objects is seen not merely as a collection of objects stored in a database, but as an entity in itself that can also be identified in the real world. Moreover, through repetitive applications of these relationships, it is possible to build hierarchies of object classes.

We define a taxonomic knowledge structure in semantic data models as a hierarchy of abstraction relationships between objects and classes of objects (see Figure 2). The horizontal plane of the taxonomy, i.e., the representation of class attributes and their values, is specified through aggregation relationships. Classification and generalization constitute the vertical relationships of the taxonomy.

Aggregation generates the construction of an aggregate object using its component parts. In Figure 2, BIRD is an aggregate object with component parts WINGS and BEAK. This is equivalent to defining wings and beak as attributes of the class birds. Classification is the elementary operation for the formation of classes. This relationship defines a class of objects as a collection of instances of objects. It links an individual object stored in the database to its immediate object class defined in the schema. In Figure 2, Tweety is classified as a canary. Generalization also contributes to the formation of classes. But it defines a class of objects as a collection of classes, not instances. Birds is a generalization of the classes of canaries, sparrows, and eagles.

ISSN: 2324-9773

Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 59-70. doi: 10.7152/acro.v1i1.12464

It should be pointed out that the use of the term *classification* is the area of semantic data models is quite different from its use in the area of classification. Using a class to describe an instance would be called *classing* or *indexing* by many researchers in the area of classification, rather than *classification*, the term used in the area of database systems. These classificationists would regard *classification* as the formation of classes, which database researchers, as just mentioned, call *generalization*.

2.2 Information Retrieval Thesauri

Information retrieval systems are used for managing collections of documents. Each document is represented by a set of terms, where a term is a word or phrase denoting a concept [Pao 89]. First, the document is analyzed in order to determine its content. Then terms which best reflect the document's content are assigned to it. This process is referred to as *indexing*. (Note: in the database area, this would be called *classification*). Assume that indexers are free to use any terms to index a collection of documents. There is no guarantee that the collection will be indexed consistently. Furthermore, searchers who use indexing terms to retrieve documents, would have to guess which terms were actually used to index the documents they with to retrieve. To avoid these problems, a thesaurus is the most commonly used device in information retrieval to provide vocabulary control [Lancaster 86]. By linking terms together, a thesaurus provides a wide range of controlled options for indexing and searching [Soergel 74].

Three basic types of term relationship are found in thesauri: synonymy, associative, and hierarchical. Figure 3 shows the typical thesaural structure, with broader terms referenced as BT, narrower terms as NT, and associative (generally related) terms as RT.

In the context of conventional thesauri, only hierarchical relations have direct implications for taxonomy. A term X is said to be broader than Y if an inclusive search of documents indexed by X also requires documents indexed by Y to be retrieved. Generic relationships BT/NT between concepts establish a hierarchy of class inclusions. Narrower concepts share the characteristics of their broader concept, plus at least one other characteristic. Thesaural hierarchies may be less strict, admitting other relationships, for example, whole-part and instance. In other words, whenever a relationship serves the retrieval function of inclusive search, it is appropriate to include it in the hierarchy [Soergel 85].

2.3 Semantic Networks in Artificial Intelligence

One of the goals of artificial intelligence is development of techniques which enable computers to perform tasks that require human intelligence when performed by humans [Tanimoto 87]. These techniques cover a broad range of issues, such as search, perception, reasoning, and learning. But perhaps one of the most important aspects of artificial intelligence is knowledge representation: how to describe the world so that computers can manipulate, and put to use, this information in an intelligent fashion. This section considers semantic networks as a formalism for representing knowledge in artificial intelligence systems.

Semantic networks represent the world (i.e., objects and relations) in terms of nodes and links. This semantic network representation was introduced by Quillian [Quillian 68], based on a

ISSN: 2324-9773

model of human memory in which concepts and relationships between concepts were represented as nodes and links, respectively. The meaning of a concept (or word) was defined by linking its corresponding node to related concepts. Such grouping of nodes resembled the organization of a dictionary [Brachman 79]. To make inferences about two concepts, the paths from their corresponding nodes had to be followed until an intersection was reached. This process was referred to as spreading activation. As a model of human memory, Quillian's work has been questioned by psychologists [Solso 88]. But his ideas gave way to a new approach for representing knowledge in computers. It should be pointed out, however, that models using explicit relational indicators between indexing terms in controlled-vocabulary indexing were developed earlier [Lancaster 72], including that of Farradane based on principles of psychology and first described in 1950 and subsequently developed further [Farradane 67]

One of the attractive features of semantic networks is their ability to represent taxonomic structures. This is possible through the creation of three kinds of links, commonly referred to as has-part, is-a, and instance-of (Figure 4). These links are directional and may be considered attributes connecting one node with a second node (the attribute value). The is-a link is an inclusion relation between classes and subclasses. The instance-of link relates instance nodes to their respective classes.

Facts about an object (i.e., node) may be inferred from other nodes to which it is linked. Inferences in semantic networks are executed by following links between nodes, as in Quillian's spreading activation process. An important type of inference in knowledge representation is property inheritance. In semantic networks properties can be inherited from a class down through its subclasses via the is-a links connecting them.

2.4 Mental Structures in Cognitive Psychology

The field of cognitive psychology is concerned with how human beings acquire and use information from the world. It deals with psychological processes like learning, pattern recognition, and memory. Two basic cognitive processes are grouping and relating information [Solso 88], essential for the formation of taxonomies.

Human memory can be viewed as a mental database. An individual acquires knowledge and keeps it in memory [Sanford 86]. There is strong evidence for the presence of hierarchical structures in memory. It has been observed, for example, that a word seems to activate not only its own representation in memory, but also the name of the category to which it belongs [Posner 73]. Classifying words (i.e., labeling them as to their meaning) takes less time if words belong to the same category rather than to a more general category, suggesting a mental hierarchy of concepts too. Another study mentioned in [Posner 73] shows that statements which are not in a hierarchical form tend to be altered in memory into a hierarchical representation.

Most psychologists believe concepts are indeed stored in some kind of hierarchical representation in human memory [Solso 88]. Concepts allow us to classify (i.e., describe in terms

ISSN: 2324-9773

Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 59-70. doi: 10.7152/acro.v1i1.12464

of classes) objects in our world. Apparently, an individual object is not perceived as a unique phenomenon but as an instance of a concept with which we are already familiar [Smith & Medin 81]. New concepts are formed by combining previous concepts into new ones. Generalization of concepts is a special kind of concept combination process. It constitutes the means by which similar concepts are grouped into a more general concept.

3. LEVELS OF ABSTRACTION IN TAXONOMY

Taxonomies inherently imply levels of abstraction. The root (top) of the taxonomy has the most general concept(s). The most specific concepts are the leaves at the bottom of the taxonomy. Each level of the taxonomy thus represents a different level of abstraction. A critical issue in the use of taxonomies thus involves the choice of an appropriate level of abstraction. This issue has been addressed in some detail in cognitive psychology and in information retrieval. Less attention has been given to levels of abstraction in semantic data models and semantic networks.

In the remainder of this section, we discuss abstraction levels in relation to creating and using taxonomy in areas considered in the preceding section. In particular, we introduce the notion of *basic levels* in cognitive science, and discuss this in relation to thesaural taxonomy. We conclude with proposals for research.

3.1 Basic Levels

Human beings are believed to organize concepts in memory hierarchically, from general to more specific. According to [Rosch 78], there exists a cognitively basic level of abstraction in these taxonomies. Basic-level concepts are often used by humans when they carry out cognitive tasks such as perception, learning, and communication.

Basic concepts seem to contain a great deal of information (see Figure 5). Members of a higher-level class share fewer attributes in common than members of a basic concept: the class of birds includes a wide variety of eagles, penguins, canaries, etc. Subclasses of basic concepts, like the class of emperor penguins, have more attributes in common among their members. But subclasses also share more attributes with each other: the class of Adelie penguins and the class of emperor penguins have many attributes in common. Thus, concepts at the basic level are clearer than, and better distinguished from, other concepts.

Obviously, objects are not always associated with their basic level. In addition, context plays a role in determining the level of abstraction at which objects will be identified. A birdwatcher will probably think about penguins in terms of emperor and Adelie species, whereas in most situations other persons will just have a general idea of what a penguin is.

The existence of basic concepts in human taxonomies causes generalization to take place at levels above the basic level and specialization at levels below it [Lakoff 87], as shown in Figure 5.

ISSN: 2324-9773

3.2 Abstraction Levels in Database Systems and Semantic Networks

Some semantic data models have distinguished situations which involve generalizations and others which involve specializations [Abiteboul & Hull 87].

Specialization implies that a class of objects can be decomposed into more specific classes. This takes place when a database designer defines potential subgroups for instances of a class. For example, a person might be a doctor, or a patient, or a nurse. Note that these subclasses may overlap. Also note that a person does not need to belong to any of these subgroups. Data models sometimes require the union of subclasses to be equal to their superclass.

Generalization, on the other hand, groups classes into more generic classes. This process implies that the original classes are in most cases disjoint. That is, a designer often imposes a condition that an instance of class belong to one and only one of its immediate subclasses. For example, cars and boats can be generalized into a vehicle class; in this specific case, the two original classes are mutually exclusive. Property inheritance is said to be upwards, in the sense that common properties of subgroups are abstracted to define their superclass.

Whether generalization or specialization takes place during the design of a database depends on which entities in the application are identified first (for example, person and car in the preceding paragraph) and how other entities are introduced later into the schema (for example, doctor and vehicle). Some models adopt either a bottom-up design by means of generalizations or a top-down approach through specializations [Hull & King 87]. Other models offer a construct for overlapping classes and another for disjoint classes, which might be interpreted as constructs for specialization and generalization, respectively.

A computational model of human memory which accounts for basic levels has been proposed by [Fisher 88]. This model is claimed to be consistent with human taxonomies of concepts. COBWEB, the system described in this study, is a conceptual clustering system which classifies objects based on their descriptions. Classification of an object goes along nodes that maximize both the sum and agreement of attribute probabilities of the object. An object is recognized at the node that is best predicted by the object's attribute values. The model is able to identify an object with respect to its appropriate basic-level concept.

3.3 Basic Levels and Choice of Thesaurus Terms in Information Retrieval

Basic levels suggest that humans often prefer concepts at a particular level of abstraction [Fisher 88]. Level of abstraction in a thesaurus determines the level of term specificity. For both indexing and searching, usage of terms depends in part on their position in the hierarchy, that is, on their level of specificity. Indexing methods usually choose indexing terms which reflect accurately the specificity of subjects covered in a document. A common indexing tenet is to index a substantively-discussed concept in a document to the most specific term available in the thesaurus.

Choice of terms used to formulate search requests also depends on the number of documents posted to a term. If a term has a large number of postings, it is expected that most documents retrieved by that term will probably not be relevant. For this reason, terms which have relatively few postings are sometimes preferred for searching documents, thereby making searches more

ISSN: 2324-9773

Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 59-70. doi: 10.7152/acro.v1i1.12464

effective.

[Weinberg & Cunningham 85] discuss some common assumptions about the relationship between the level of specificity and the number of postings. It is widely assumed that general terms (terms at higher levels) will have more postings than specific terms (terms at lower levels). Terms that have recently been added to a thesaurus tend to have fewer postings, and they usually are of a more specific nature. However, in the case of a thesaurus or a collection of documents which focuses on a particular discipline, the number of postings for a term would depend on whether the term is central or peripheral to the discipline and on its level of specificity. It is believed that specific terms will be more often used to index central topics, and general terms will be used more for peripheral topics.

In Table 1 we show the relationship between term specificity and postings for the MeSH tree structure corresponding to Endocrine Diseases. Level 1 corresponds to the broadest term, i.e., Endocrine Diseases, the second level to its immediate narrower terms, and so on. Hence, we would expect more specific terms to have larger number of postings. For this Endocrine Diseases structure, this correlation was equal to -0.20422, with a significance level of 0.02185 [Weinberg & Cunningham 85, p. 368]. Level 3 contains the highest number of terms and postings. This level, quite interestingly, corresponds to the middle level of the hierarchy. Perhaps this is because most terms at this level correspond to fundamental concepts with respect to endocrine diseases. This empirical evidence, of greater number of postings at intermediate levels, seems to contradict the common beliefs mentioned earlier.

Level	Terms in Level	Sum of Postings	Postings Mean	Postings Median
1	1	3134	3134	3134
2	12	47280	3940	1133
3	51	132055	2589	1348
4	35	51004	1457	822
5	2	1408	704	704

Table 1. Postings statistics for *Endocrine Diseases* structure from MeSH (postings data from [Weinberg & Cunningham 85]).

4. PROPOSALS FOR RESEARCH ON ABSTRACTION LEVELS IN TAXONOMIES

There are many potential areas for future research on levels of abstraction in taxonomies, both in general and in specific application areas. Some descriptive work is needed, in order to investigate the structures of existing taxonomies. In addition, there are several issues related to the design and use of taxonomies which can be investigated.

Further work should be done on the relationship between the level of abstraction and the number of postings in thesauri in order to investigate the hypothesis that most postings will occur at intermediate levels. It would also be interesting to investigate the levels at which terms are added or deleted. Similar questions could be asked in other areas.

ISSN: 2324-9773

Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 59-70. doi: 10.7152/acro.v1i1.12464

The sizes of taxonomic structures in different applications can be compared. The size could be measured in terms of the number of levels, the number of concepts, and the branching factors. Many information retrieval thesauri are available for study. It might be more difficult to obtain data in other areas; however, approximate measures could be used.

We have considered here only the static properties of taxonomies. However, most taxonomies support at least a limited set of operations, and some include procedural information. There is growing interest in object-oriented systems, and inheritance is a critical feature of such systems. It might be useful to consider the relationships between levels of abstraction and inheritance and storage of procedural information.

REFERENCES

Abiteboul, S., Hutt, R. "IFO: a formal semantic database model." **ACM TODS** 12 (4), Dec. 1987, 525-565.

Brachman, R.J. "On the epistemological status of semantic networks." In N. V. Findler (ed.), **Associative networks: representation and use of knowledge by computers**, Academic Press, New York, 1979, 3-50.

Brodie, M. L., Manola, F. "Database management: a survey." In J. Myopolous and M. Brodie (eds.), **Readings in artificial intelligence and database systems**, Morgan Kaufmann Publishers, Inc., San Mateo, California, 1988, 10-34. Also in J. W. Schmidt and C. Thanos, **Fundamentals of knowledge based systems**, Springer-Verlag, New York, 1988.

Brodie, M. L., Ridjanovic, D. "On the design and specification of database transactions." In M. L. Brodie, J. Myopolous and J. W. Schmidt (eds.), **On conceptual modeling: perspectives from artificial intelligence, databases, and programming languages**, Springer-Verlag, New York, 1984, 277-306.

Evens, M. W., Litowitz, B. E., Markowitz, J. A., Smith, R. N. and Werner, O. **Lexical-semantic relations: a comparative survey**. Linguistic Research, Inc. Edmonton, Alberta, Canada. 1983.

Farradane, J. "Concept organization for information retrieval." **Information Storage and Retrieval**, 3(4), 1967, 297-314.

Fisher, D. H. "A computational account of basic level and typicality effects", **Proc of the 7th Nat. Conf on Artificial Intelligence**, St. Paul, Minnesota, Aug. 21-26, 1988, Vol. 1, 233-238.

Furtado, A. L., Neuhold, E. J. **Formal techniques for database design**. Springer-Verlag, Berlin, 1986.

Hawryszkiewycz, I. T. **Database analysis and design**. Science Research Associates, Inc. Chicago. 1984.

Hull, R., King, R. "Semantic database modeling: survey, applications, and research issues." **ACM Computing Surveys**, 19 (3), Sep. 1987, 201-260.

Lakoff, G. **Women, fire and dangerous things: what categories reveal about the mind.** University of Chicago Press. Chicago, Illinois. 1987.

ISSN: 2324-9773

Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. Proceedings of the 1st ASIS SIG/CR Classification Research Workshop, 59-70. doi: 10.7152/acro.v1i1.12464

Lancaster, F. W. **Vocabulary control for information retrieval.** 1st ed. Information Resources Press. Arlington, Virginia. 1972.

Lancaster, F. W. **Vocabulary control for information retrieval. 2nd ed. Information Resources Press.** Arlington, Virginia. 1986.

Pao, M. L. **Concepts of information retrieval.** Libraries Unlimited, Inc. Englewood, Colorado. 1989.

Posner, M. I. **Cognition: an introduction.** Scott, Foresman and Co. Glenview, Illinois. 1973.

Quillian, M. R. "Semantic memory." In M. Minsky (ed.), **Semantic information processing,** MIT Press, Cambridge, Massachusetts, 1968, 227-270.

Rosch, E. "Principles of categorization." In E. Rosch and B. B. Lloyds (eds.), **Cognition and categorization,** Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1978, 27-48.

Sanford, A. J. **Cognition and cognitive psychology.** Weidenfeld and Nicolson. London. 1986.

Smith, E. E., Medin, D. L. **Categories and concepts.** Harvard University Press. Cambridge, Massachusetts. 1981.

Soergel, D. **Indexing languages and thesauri: construction and maintenance.** Melville Publishing Co. Los Angeles, California. 1974.

Soergel, D. **Organizing information: principles of data base and retrieval systems.** Academic Press, Orlando, Florida. 1985.

Solso, R. L. **Cognitive psychology.** 2nd ed. Allyn and Bacon, Inc. Boston. 1988.

Tanimoto, S. L. **The elements of artificial intelligence.** Computer Science Press, Inc. Rockville, Maryland. 1987.

Weinberg, B. H. "Interactions of statistical phenomena in human and automatic indexing." **Proc of the 45th Annual Meeting of the American Society for Information Science,** Columbus, Ohio, Oct. 17-21, 1982, 324-326.

Weinberg, B. H., Cunningham, J. A. "The relationship between term specificity in MeSH and online postings in MEDLINE." **Bulletin of the Medical Library Association,** 73 (4) Oct. 1985, 365-372.

ISSN: 2324-9773

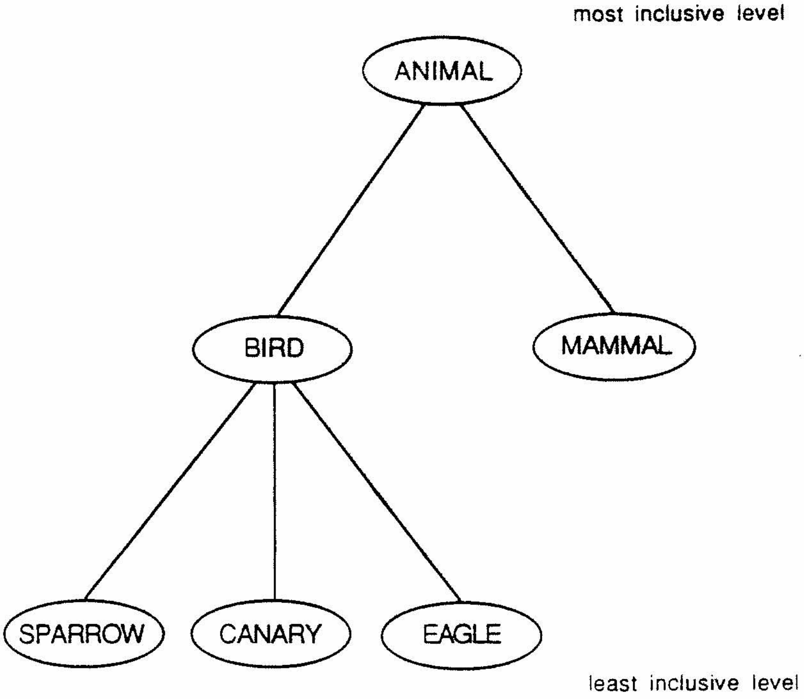

Figure 1. A simple taxonomic structure.

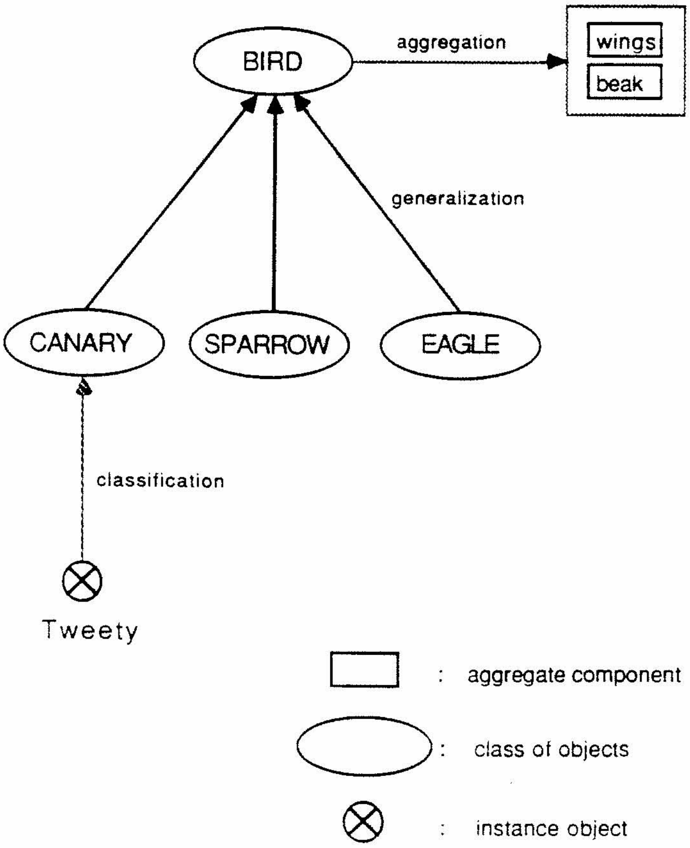

Figure 2. Abstraction relationships in semantic data models.

ISSN: 2324-9773

BIRDS

NT CANARY
EAGLE
SPARROW

BT ANIMAL

RT BEAK
ORNITHOLOGY
WINGS

Figure 3. Term relationships in thesauri.

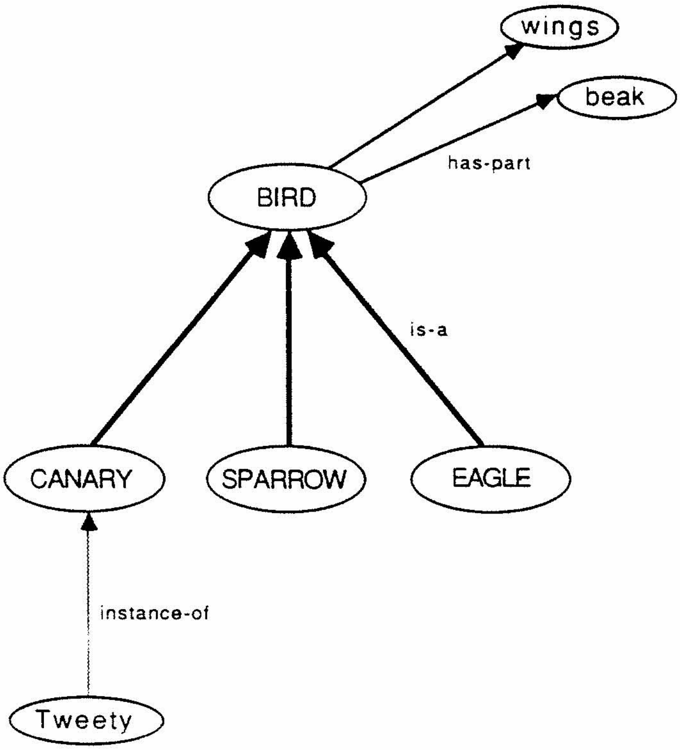

Figure 4. A taxonomic structure in semantic networks.

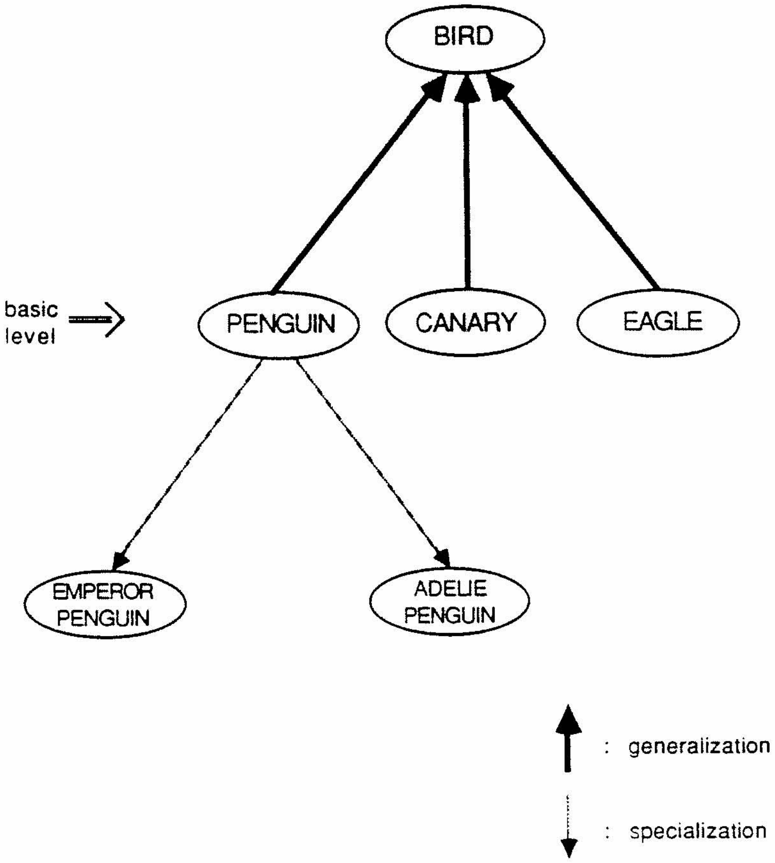

Figure 5. Basic levels of taxonomies.

ISSN: 2324-9773