Dimitris A. Dervos[1], and Anita Coleman[2]
[1]Information Technology Dept., Alexander Technology Educational Institute, Thessaloniki, Greece. Email: *dad@it.teithe.gr*
[2]School of Information Resources & Library Science, University of Arizona, Tucson, USA. Email: *asc@u.arizona.edu*

This is a preprint. Published. Dervos, D. and Coleman, A. (2006). A Common Sense Approach to Defining Data, Information and Metadata. Proceedings of the Ninth International Society for Knowledge Organization 2006 Conference, Vienna. June 2006, Edited by C. Swertz. Berlin: Ergon.

A Common Sense Approach to Defining Data, Information, and Metadata

Abstract: Many competing definitions for the terms data, information, metadata, and knowledge can be traced in the library and information science literature. The lack of a clear consensus in the way reference is made to the corresponding fundamental concepts is intensified if one considers additional disciplinary perspectives, e.g. database technology, data mining, etc. In the present paper, we use a common sense approach borrowed from the data mining community, which has successfully solved many data processing problems, to selectively survey the literature, and define these terms in a way that can advance the interdisciplinary development of information systems.

1. Introduction

Many competing definitions for the terms data, information, metadata, and knowledge can be traced in the library and information science literature (Faradane, 1979; Buckland, 1991, McCrank, 2002). The *ALA Glossary of Library Terms* (1943) does not mention these three terms although the *Online Dictionary of Library and Information Science* (2005) modeled on the *ALA Glossary of Library and Information Science* (1983) does. A closer look at the definitions reveals that they appear to differ from the way they are agreed upon and used in the traditional computer science, database technology, and data mining communities. Similarly, we find metadata is another term that differs in the library and information science literature from that of knowledge management.

In this paper, we assume that competing and divergent definitions get in the way of true interdisciplinary collaboration between computer scientists, library/information scientists, and all those who need to work together to solve problems. They not only prevent the accurate accumulation of data, information, or knowledge, but also hinder the development of systems that can truly move us into the era of the information society. Since many disciplines study information today, information, data, metadata, and knowledge also have current and emergent meanings. Definitions for these fundamental concepts in the many academic disciplines that study information and its variants (data, knowledge, wisdom, beliefs, etc.) however, continue to be widely divergent (Debons et al. 2005).

The common sense approach established is outlined in Section 2. In Sections 3, 4, and 5 we define the concepts of data, information, and metadata, respectively. Next, we consider a number of interesting implications in Section 6, and conclude in Section 7.

2. The Common Sense Approach

Debons (2005a) proposed that the basic notions of data, information, and knowledge can be defined by observing the following two preconditions (Figure 1):

PRECONDITION-1:
A "living species" approach is adopted, i.e. social/organizational systems are not addressed, at this stage, the focus being on the individual living organism (the human being comprising only one case of the many), and technology is ignored.

PRECONDITION-2:
Reasoning builds upon a finite number of simple assumptions made initially.

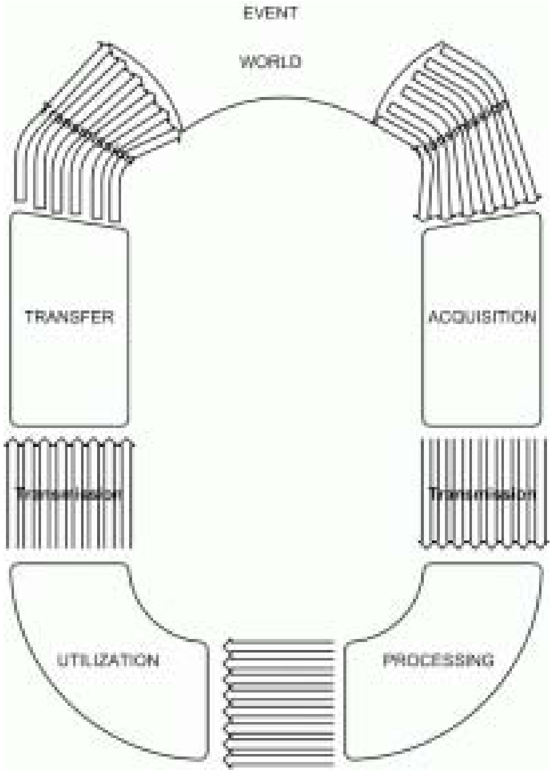

Figure 1 A living species interaction with/understanding of the environment: (Debons 2005a)

In compliance with preconditions -1, and -2, the following two assumptions can be made:

ASSUMPTION-1:
Information is at a higher level than data

ASSUMPTION-2:
Knowledge is at a higher level than information

3. Data

McCrank (2002, p. 627) defines data as follows: “data are what is given in the smallest units, from digits to arrays and points to lines, and bits of information which are encountered, collected, or inferred, and manufactured, that are neither facts nor constitute evidence by themselves. These are the raw material for building information.” McCrank further defines facts and this definition provides a clue as to how disciplinary differences can sometimes be overcome. “**Facts** are things done, that is deeds or acts made into something known from (from *facer*, to make, so something made), which have had or do have actual existence, and are true and pertain to objective reality…facts are supposedly stable,actual, and real and can therefore be made evident (they are represented but must be presented), Similarly, for those in the data mining field, data are distinguished from facts. “A fact is a simple statement of truth” (Roiger and Geatz, 2003, p. 5). Observing the world, the most primitive type of autonomous intellectual activity one can think of is the recording of facts that relate to the object/event upon which observation is focused. Facts can only be recorded (subsequently: processed) once they are represented properly in the appropriate model space; that is when data come into play.

DEFINITION:
Data represent real world facts.

For example, data may comprise the outcome of measurements conducted in relation with real world phenomena (e.g. rainfall values for a given set of geographical locations over a period of time). Also, data may relate to values of attributes that characterize entities and/or relationships between entities in real world application model situations (Ramakrishnan 2004).

4. Information

Faradane (1979) defined information as any physical form of representation, or surrogate, of knowledge, or of a particular thought, used for communication. A little less than twelve years later, Buckland (1991) refined information further and distinguished between information-as-thing, information-as-process, and information-as-knowledge. He convincingly argued that information-as-thing is what is dealt with in modern day information systems. Machlup (1980), one of the earliest researchers who tried to measure the information economy defined knowledge as content and information as process. However, his initial attempts to provides proofs for these definitions were limited to measurement of the activities of scientists and researchers, and ignored the processing work done by librarians and records managers. Later, Machlup and Mansfield noted, “Information is not just one thing. It means different things to those who expound its characteristics, properties, elements, techniques, functions, dimensions, and connections.” (1983, p. 4).

Thus, information is hard to define, directly. However, it is felt to relate to the communication and interpretation of data, the process of “becoming informed”, and “informing”. Interestingly enough, despite lacking a direct formal definition, the concept is better understood by the influence carriers that are said to be information rich have on the environment.This, information may be defined as follows:

DEFINITION:

Information is revealed each time data are interpreted successfully in the direction of increasing benefit, profit, or pleasure, as the latter are realized by some intellectual activity.

The human mind appears to favor savings on the overhead associated to processing data for the purpose of extracting useful information from them. In this respect, shortcuts that interpret data in a most concise and comprehensible way are taken to comprise *clever things*. For example, a plot that occupies one third of an A4 page may comprise a shortcut to two A4 pages of tabular data in revealing the same information on how, say, a dependent variable responds to the way values are assigned to an independent variable. A one-page entity-relationship (ER) diagram comprises a shortcut revealing information in relation to an application model the textual description of which may utilize, say, five A4 pages (Chen 1976). A rule, like *heavy smokers have a high probability to develop lung cancer* comprises a shortcut to interpreting data of thousands patient records.

Shortcuts of the type described in the previous paragraph are felt to be utilized in the intellectual activity that compiles/accumulates knowledge (Miller, 1956). As it has been stated already, the latter is assumed to be one level up in the intellectual hierarchy chain when compared to information.

Most people will agree that information is certainly something that helps produce knowledge. Additionally, the truth-value of information can be an important criterion in both the determination of information itself as well as in measuring it. Truth-value from mathematical logic may be a better criterion than accuracy[1], because truth-values can be calculated based on two values only, true or false. No further attempt is made in the direction of realizing and defining the concept of knowledge at this stage.

5. Metadata

Metadata are often glibly and ambiguously defined as “data about data”. A somewhat more involved explanation is useful before we define metadata. To facilitate the processing stage, data need be organized in structures that group values in accordance with the semantics of the facts they relate to.

DEFINITION:
Metadata are tags/labels assigned to data instances and structured in order to make them comprehensible and/or facilitate the processing that extracts information from data corpora.

For example, when observation targets the academic performance of students in an academic establishment, the corresponding set of metadata could include labels like: *Student ID, Department, Year of Entry, Course ID, Grade*, etc. In the case of an information resource or information package, metadata labels could include *Format, Form, Creator, Title,* etc. When the target model involves conceptual ideas expressed through language only, metadata labels that attempt to categorize and capture both data as well as information could include *Process, Object, Phenomenon,* etc.

6. Implications/Discussion

[1] Accuracy is often only one criterion in multi-dimensional and complex models of information and data quality (Olaisen, 1990, Fox et al, 2004).

Debons' approach to defining data, information, metadata, and knowledge may be extended in the direction of adopting a human-centric approach in order to identify the current stage of the forefront in human civilization. Everyone agrees that it is post-agricultural, and post-industrial. Interesting issues to argue upon relate to questions like:

- *The information age: Has it begun? Is it nearing? Is it here?*
- *The knowledge age: Has it begun? Is it nearing? Is it here?*
- *What is it that comes after the knowledge age? The age of wisdom?*

Confusion begins to develop when software product vendors use exotic terms to name their products for marketing purposes. How can one compete in the information management market today when IBM have been marketing their premier transactional and hierarchical database management product under the name of IMS (Information Management System) since the early 60's (IBM 2006)?

Once again, one needs to rely on common sense in order to find a way out of such a maze:

ASSUMPTION:
Major transitions in human civilization relate to qualitative changes in the way people perceive the world and function in relation to the (external) environment in everyday life, not to quantitative ones.

Applying this to the case of data and data processing, one notes that:

1. The concept of data has been defined and it is fully understood
2. The concept of data has been quantified. The storage space a data corpus occupies remains invariant upon transportation from system to system, provided that the technology that materializes the representation remains invariant
3. When it comes to storing data on digital media, the unit that measures data 'quantity' is well defined: the bit.
4. In the developed part of the world today, everyday human activity is shaped, to a great extent, by the data storing and data processing operations of digital devices: one considers a digital organizer as an extension to his/her memory – for as long as an instance of data is registered with the device, s/he no longer cares about remembering it; an alarm sound will go on when time comes for that telephone call to be made (the name and the phone number of the other party flashing on the screen of the organizer).

Considering the above 1-4 criteria in order to identify whether information has come of age, one notes that:

1. Today, the concept of information is realized only indirectly, not directly. This is analogous in a way to the case of elementary particles in Physics: one cannot measure their properties directly (i.e. mass/energy equivalent, wavelength, electric charge, etc.), but only indirectly via the influence they have to their environment.
2. The concept of information has not yet been quantified. The information inherent to a given dataset can usually be extracted in a most concise way (i.e. a shortcut) via, say, a

graph, a rule, or (even) a pattern. One is not sure of having achieved the most concise way of extracting a given instance of information until (probably) a better shortcut is invented

3. There exists no scheme for measuring information (or a way to model it). For example, the present situation is far from having one, for example, claim that "John called me last night and in five minutes of talk he revealed five *infotrons* worth of information to me, whereas Mary called a bit later and in just three minutes she revealed ten *infotrons* worth of information on the same topic".
4. Human activity is far from utilizing technology that incorporates information processing in everyday life, today. Many visions such as Bush's Memex (**Bush, 1945**) are indicative of how things will be when information processing is to become a routine of everyday human life, namely when technology: (a) models the user profile/interests/preferences, (b) senses the current context of the human, (c) processes information relevant to the previous a and b, and (d) interacts with the human by presenting the information in a subtle, and non-intrusive way. Obviously, the human has some way to go before s/he reaches the point where his/her everyday activities are shaped to co-exist and co-function harmonically with technology in such a mode.

Considering the above:

COLLORARY-1:
Information remains to be controlled, quantified, modeled, and to be fully understood as a concept
COROLLARY-2:
The forefront of our civilization, with regard to the technology advances made and the way those advances shape the everyday life of humans, is still at the data processing age. Additional time and research effort still need to be invested, until the information society comes of age.

Humans make associations and relationships between concepts and ideas all the time. Khoo and Na (2005) reviewing semantic relations research conclude that while they are important in information processing applications and point the way forward for information science research, the precise application and uses of fine-grained semantic relations are as yet known. We suggest a specific use. Collocating the definitions for the fundamental terms of data, information, knowledge, and metadata, from the various disciplines, such that the various groups that work with them are aware of differences is a first step towards information transformation systems (Neelameghan, 1972). By collocation is also meant the process of bringing together to form a junction (among subjects or disciplines), such that explicit and detailed semantic and others usch as case relations and lexical-semantic relationsrelations can be further identified in various contexts, in order to place them in multiple logical orders or for using them in precise measurement of the abstract concept/idea :

7. Conclusion

Machlup and Mansfield suggested, as early as 1983, that "most of the confusion caused by the use of the term information science in its broadest sense could be avoided by the addition of the plurals". That is, many disciplines comprise the information sciences, like the social sciences and natural sciences. While writings that have examined information and related phenomena are not exactly unique (Cana, 2003), the imperative of the activity-theoretical approach to information science (Hjorland, 1997), interdisciplinarity, multi-disciplinarity and the “viewpoint warrant” (Beghtol, 1988) suggest that we first identify the terms and their competing definitions from the many branches of knowledge, and then work consensually towards acceptance of the fundamental ones such that they are sharable and applicable across interdisciplinary domains.

In the present paper, we have used the common sense approach to show how some definitions can be developed in a way that promotes the usage of a common vocabulary in many disciplines. We hope that our proposal can be fruitful in interdisciplinary work, thereby, leading to the growth of the information sciences and the development of information systems that truly make possible the realization of the information society.

Acknowledgement

The authors are thankful to Richard J. Hartley for his prompt comments on various drafts of the present paper.

References

Beall, J. (2006). Metadata and Data Quality: Problems in the DigitaL Library. Journal of Digital Information 6 (3). Retrieved 27 February 2006. http://jodi.tamu.edu/Articles/v06/i03/Beall/

Beghtol, C. (1988). General Classification Systems: Structural Principles for Multidisciplinary Specification. *Structures and relations in knowledge organization: Proceedings of the 5th International ISKO Conference, Lille, 25-29 August 1998;* eds. W. Mustafa el Hadi, J. Maniez, S. Pollitt. (Advances in Knowledge Organization - 6). Würzburg: Ergon, 1998, pp. 89-96.

Buckland, M. (1991). Information as Thing. *Journal of the American Society for Information Science* 42 (5): 351-360.

Bush, V. (1945). As We May Think. Atlantic Monthly. Retrieved 24 February, 2006. http://www.theatlantic.com/doc/194507/bush

Cana, Mentor (2003). The Understanding of Information and Information Science. Retrieved 4 November, 2005. http://www.kmentor.com/socio-tech-info/archives/000050.html

Chen, P.P. (1976). The Entity-Relationship Model – Toward a Unified View of Data. *ACM Transactions on Database Systems*, 1(1):9-36.

Debons, A., Zins C., Beghtol C., Harmon G., Hawkins D., Froehlich T.J. et al. (2005). T*he Knowledge Map of Information Science*, 2005 ASIS&T Annual Meeting, October 28 – November 2, Charlotte N.C., U.S.A.

Faradane, J. (1979) The Nature of Information. *Journal of Information Science* 1: 13-17.

Fox, C. Levitin, A, and Redman, T.C. 1996. Data and data quality. (pp. 100-122). In *Encyclopedia of Library and Information Science*. N.Y.: Marcel Dekker.

Hjorland, B. (1997). Information Seeking and Subject Representation: An Activity Theoretical Approach to Information Science. Westport, Conn.: Greenwood.

IBM (2006). The IMS Family. Retrieved 1 February 2006. http://www-306.ibm.com/software/data/ims/

Khoo, C. S.G. and Na, J. (2005). Semantic Relations in Information Science. *Annual Review of Information Science and Technology* 40. Medford, N.J.: Information Today.

Machlup, Fritz. (1980) *Knowledge: Its Creation, Distribution, and Economic Significance. Vol 1, Knowledge and Knowledge Production*. Princeton, NJ: Princeton University Press.

Machlup, Fritz, and Mansfield, Una. (1983). Editors. The Study of Information: Interdisciplinary Messages. New York: Wiley.

McCrank, Lawrence. (2002). Historical Information Science. Medford, N.J.: Information Today.

Neelameghan, A. (1972). Systems Approach in the Study of the Attributes of the Universe of Subjects. *Library Science with a Slant to Documentation* 9 (4): 445-472.

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63: 81-97. Retrieved 27 February 2006. http://www.well.com/user/smalin/miller.html

Olaisen, Johan. "Information Quality Factors and the Cognitive Authority of Electronic Information." In Information Quality: Definitions and Dimensions. I. Wormell (ed.). London: Taylor Graham, 1990, pp. 91-121.

Online Dictionary of Library and Information Science. Retrieved 4 November, 2005. http://lu.com/odlis/odlis_s.cfm

Ramakrishnan R., and Gehrke J. (2004). *Database Management Systems*, 3rd Edition, McGraw-Hill Science/Engineering/Math

Roiger R., and Geatz M. (2003). *Data Mining: A Tutorial-Based Primer*. Boston, Mass.: Addison-Wesley Publishing.