WP. 12
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (ii): Metadata interchange

THE NATURE OF DATA [1]

Contributed Paper

Submitted by the Bureau of Labor Statistics[2], United States

I. INTRODUCTION

1. The term data is used everywhere in the statistical and computer science literature. Almost never do you see the concept defined. The apparent reason for this is that data as a term is thought to have a well-defined concept behind it. And this is not unique, as many terms are so commonly used, they cease requiring any kind of definition or justification to their use. The writer assumes everyone knows what the terms mean. <u>Data</u>, <u>function</u>, <u>interface</u>, <u>object</u>, and <u>service</u> are just a few common examples in the computer science field.

2. Occasionally, we run into definitions of commonly used terms such as data. We will look at two of them in this paper. The fundamental question is whether these definitions describe the essence of the concept. What we want to know from any definition is the following: What <u>is</u> the thing we are trying to describe? These prior definitions of data do not answer this question, as we shall see. Once you start digging below the surface, the lack of good definitions leads to confusion. This paper is an attempt to shed light on these and related issues.

---

[1] The opinions expressed in this paper are those of the authors and do not necessarily reflect the official policies of the Bureau of Labor Statistics or Farance, Inc..

[2] Prepared by Daniel W. Gillman at the US Bureau of Labor Statistics and Frank Farance at Farance, Inc, New York, NY USA. Contacts gillman.daniel@bls.gov and frank@farance.com.

3. Terminology theory is the study of concepts and their representations in special languages. It focuses on the essential characteristics of concepts, therefore on what a concept is. Applying the theory, we define data in a new way, by defining what data is, by investigating its essential characteristics. This, in turn, provides a way to distinguish between data and information usefully. The role of metadata is clearly defined. Some other related topics are discussed as well.

4. The method for uncovering the essential characteristics of a concept is through concept analysis. See Frankfurt (2005) for a humorous but effective example. Here, we try to uncover the essential characteristics of the concept under review. Of course, a concept is described by a definition, and one that describes the essential characteristics of the concept is an intensional definition. So, our goal is to arrive at intensional definitions.

5. The paper is divided into several sections. First, we describe the essential ideas in the theory of terminology. Then, through a concept analysis, including a critique of the two definitions of data, we arrive at a new definition. From there, new definitions of information, metadata, and data element are derived. Along the way, some surprises are uncovered.

II. THEORY OF TERMINOLOGY

6. Terminology is the study of concepts and their representations in special language. It is multidisciplinary, drawing support from many areas including logic, epistemology, philosophy of science, cognitive science, information science, and linguistics. Work in the area dates all the way back to the ancient Greek philosophers.

7. To begin, we describe some useful constructs from the theory of terminology. These come from several sources (Sager, 1990; ISO, 1999; ISO, 2000). The constructs and their definitions follow below:

- **characteristic** - abstraction of a *property* of a set of *objects*
- **concept** - unit of knowledge created by a unique combination of *characteristics*
- **concept system** - set of concepts structured according to the relations among them
- **definition** - expression of a *concept* through natural language, which specifies a unique *intension* and *extension*
- **delimiting characteristic** - *essential characteristic* used for distinguishing a *concept* from related *concepts*
- **designation** - representation of a *concept* by a *sign*, which denotes it
- **essential characteristic** - *characteristic* which is indispensable to understanding a *concept*
- **extension** - set of *objects* to which a *concept* refers
- **general concept** - *concept* with two or more objects that correspond to it (e.g., planet, tower)
- **generic concept** - *concept* in *generic relation* to another that has the narrower *intension*
- **generic relation** - relation between two *concepts* where the *intension* of one of the *concepts* includes that of the other *concept* and at least one additional *delimiting characteristic*
- **individual concept** - *concept with one object* that corresponds to it (e.g., Saturn, Eiffel Tower)
- **intension** - sum of *characteristics* that constitute a *concept*
- **object** - something conceivable or perceivable

- **<u>property</u>** - observation, used to describe or distinguish an *<u>object</u>* (e.g., "Dan has blue-gray eyes" means "blue-gray eyes" is the property of Dan. It is abstracted to a characteristic, color of eyes, of people - see *<u>characteristic</u>.)*
- **<u>sign</u>** - *<u>general concept</u>* whose *<u>extension</u>* contains only perceivable *<u>objects</u>*
- **<u>specific concept</u>** - *<u>concept</u>* in *<u>generic relation</u>* to another that has the broader *<u>intension</u>*
- **<u>subject field</u>** - field of special knowledge

8. Designations come in three types: A term is a verbal designation of a general concept; an appellation is a verbal designation of an individual concept; and a symbol is any other designation. Signs, through which designations are represented, are left undefined, but a sign is what a person perceives and interprets as designating some concept. Basically, however, a sign is a concept whose extension is a set of perceptible objects. Examples of signs are each of the lines and dots on this page we interpret as words, letters, and punctuation. So, what we see and interpret is not really a sign, but an object in the extension of the sign. The objects **F** and **F** are in the extension of the same sign.

9. Characteristics are used in concept formation. They are abstracted from properties of objects and are used to form the intension of concepts. The objects whose properties are abstracted into the characteristics that form the intension of some concept make up its extension. Characteristics may be concepts in their own right, too. They are used in concept analysis, concept modeling, formulation of definitions, and even term formation.

10. The term *specialization* is often used to denote the creation of a specific concept in generic relation to a given, generic, one. Examples are a dog is a specialization of a mammal and a triangle is a specialization of a polygon.

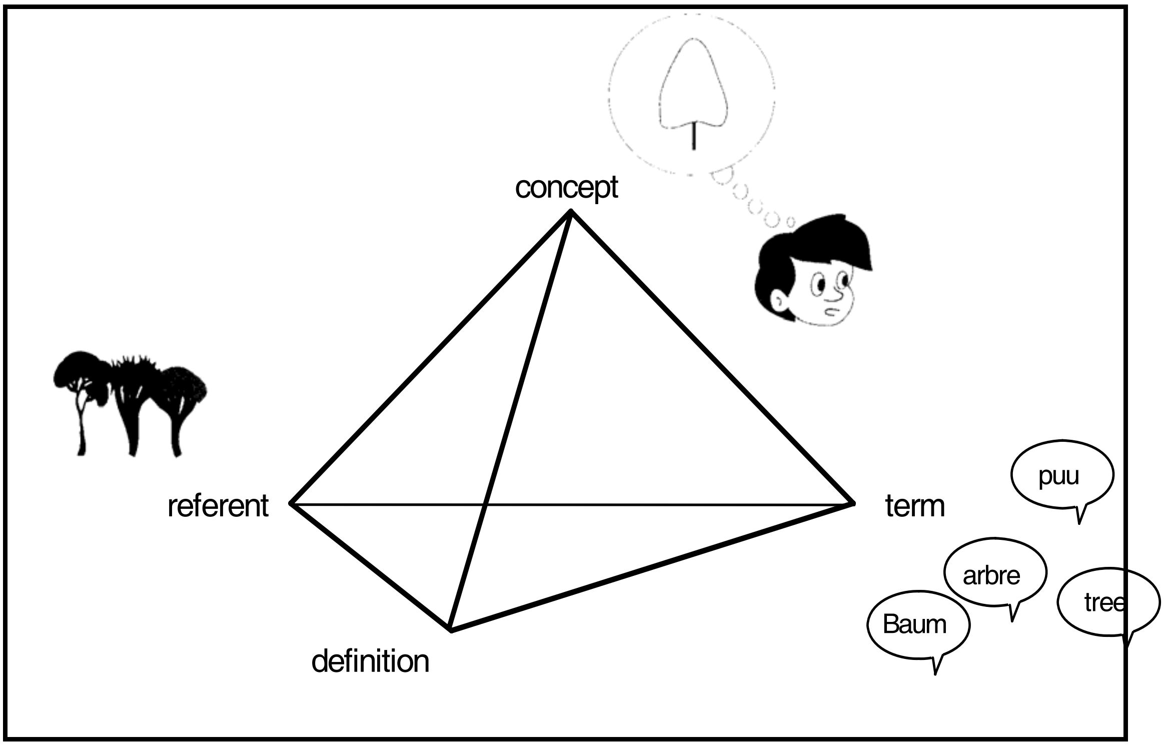

Figure 1: Relationships between referents (objects), concepts, terms (more generally designations), and definitions.

11. The ancient Greek philosophers began the study of terminology and concept formation in language (Wedberg, 1982), and they discovered a useful relationship between designation, concept, object, and definition, that is illustrated in Figure 1 (CEN, 1995). This diagram, minus the definition part, is often referred to as Ogden's Triangle (Ogden and Richard, 1989).

12. Figure 1 shows how terms, concepts, objects, and definitions are related. From the definitions above and Figure 1, several important observations need to be made:

- Concepts, terms, and objects are not the same things
- For any concept, there may be many designations (synonyms)
- For any concept, there are one or more objects in its extension
- For any concept, there may be more that one definition (especially in multiple languages)
- For each term, more than one concept may be designated (homographs)

13. Concepts are human constructions (Lakoff, 2002). No matter how well we define a concept, a complete description is often impossible. Identifying the relevant characteristics is culturally dependent. So, some objects in the extension of a concept, called prototypes, fit the characteristics better than others (Lakoff, 2002). For example, a robin fits more of the characteristics of a bird than a penguin does.

III. Data

A. Problem Description

14. Data is a concept that seems both intuitive and obvious, yet when we try to write down a definition, we struggle to avoid the circular definitions that define "data" using the term "information" and then defining the term "information" using the term "data".

15. Let us consider a definition for data from ISO (1993) The definition reads "Reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing." The main problem with this definition is that it tells the reader what data does, not what it is. Data can be reinterpreted; they are suitable for communication, interpretation, and processing. These are all things we want data to do. It is hard to argue with this, except that it is very difficult to tell whether any given objects are data or not. This is because we still don't know what data is from this definition. For this reason, an intensional definition is better - it describes what data are from the point of view of the essential characteristics of the concept.

16. Another approach is to use the familiar triangle that is divided into 4 layers: data (bottom), information, knowledge, and wisdom (top), see Figure 2 - the knowledge hierarchy (Bradley, 2004). We are unsure of precisely what this triangle conveys. The layering might convey composition, definition, or representation. For example, information could be composed of, defined by, or represented as data. In fact, Larry English (1999) defines each layer of the triangle in terms of a function on the thing in the layer below. So, information is a function on data (and some other things). However, the functions are not very precisely defined. Alternatively, the layering might convey something else.

17. English does provide a definition of data: Data is the representation of facts about things. This is required since there is no lower tier on the triangle. On the surface, this meets our criteria. It states what data is. However, on closer examination, there are several questions. Namely, what are meant by representation, facts, and things? These aren't defined, so the definition is not really intensional. For instance, is an idea a thing? What constitutes a

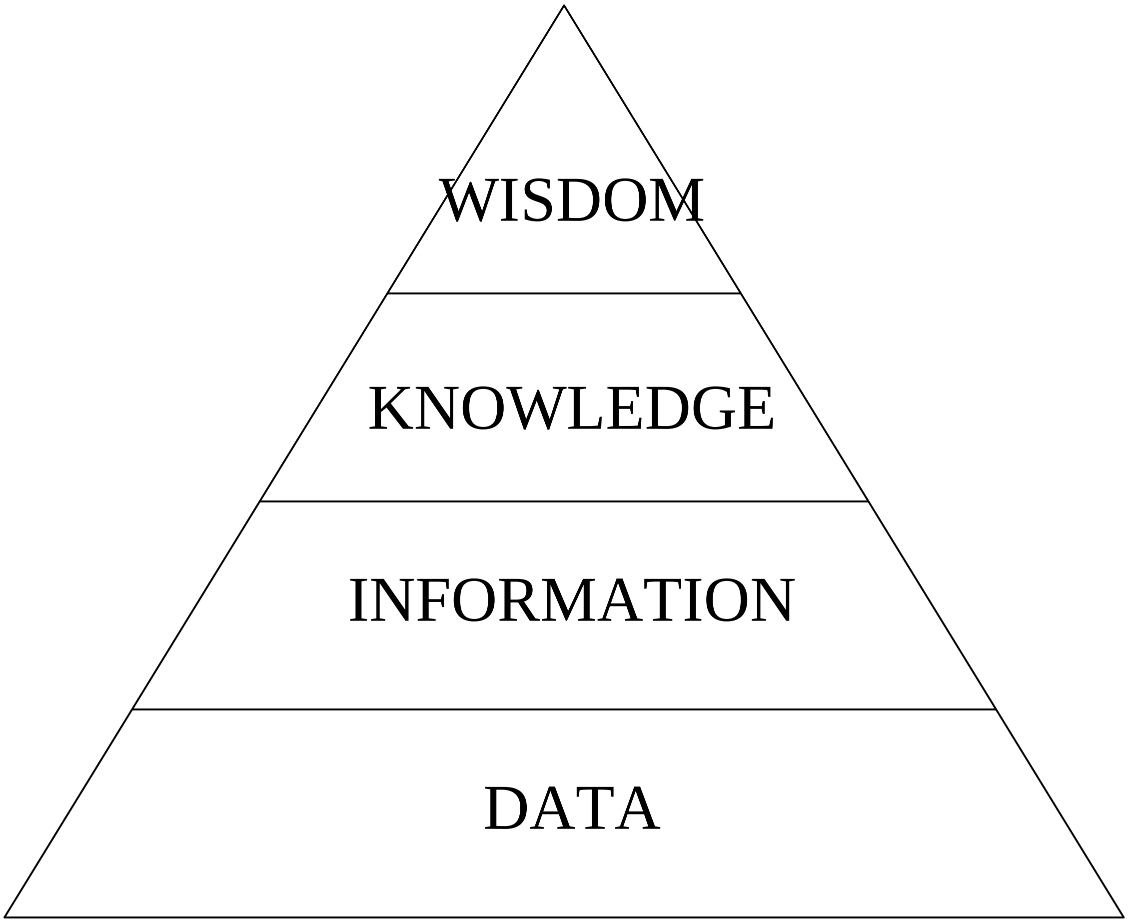

Figure 2: Knowledge Hierarchy

fact? Do data have to represent something that is true? What constitutes a representation? Do all data correspond to characters of some kind? We don't know. Therefore, we don't know the essential characteristics.

18. We believe there is a parallel relationship between data and information, i.e., neither is below the other. We aren't going to try defining knowledge and wisdom, but they might be the subject of some future paper.

19. As stated above, we believe information is a concept related to the concept of data. The topic of "information" has been explored by Langefors (1995), Shannon (1948), and others. Our presentation of data and information is consistent with Shannon's concept of information, and we include a discussion of a close association with Langefor's theory. Yet, we frame the discussion in terminology theory.

20. In this paper, rather than speak of data, we will talk about a datum and derive a definition of datum based on terminological principles. If we look at a datum in a table, say, we recognize that there are underlying concepts the datum represents. The title, column and row headers, and any footnotes in the table tell us much about our datum. These words represent concepts, and these concepts tell us how to interpret the datum. Maybe these words are not enough for a full understanding of what the datum means, however, they serve a purpose. They tell us that a datum is some designation (amount, count, code, etc.) representing a concept in a concept system.

21. An example is the US unemployment rate. If one saw the sign "4.7" in a table showing US unemployment rates for the past 12 months under the column marked "January 2006 in percent", then the reader may infer that the US unemployment rate for January 2006 was 4.7%. Knowing what "unemployment rate" means and relating it to the month of January 2006 gives the reader much information. So, again, the datum is the designation of some concept within a concept system. In order to fully explain these ideas in what follows, we will make this precise.

B. Equality on Concepts

22. A fundamental characteristic of a datum is that it can be compared with other datums[3]. So, there must be a notion of equality associated with them. A datum represents concepts, so equality on datums requires defining equality for concepts. Therefore,

Definition: A concept has **equality** if its extension is determined consistently

23. Many times, two instances of a concept don't have the same extension. Take the concept nearness, defined as proximity between pairs of objects. The extension contains pairs of objects, which are deemed near. Frank may decide that a table and chair in his room are near to each other. Dan, on the other hand, walks into the same room and decides differently. The table and chair are not near to each other; they are just too far apart for his liking. Frank's and Dan's extensions are not consistently determined, they differ, so their concepts of nearness are not equal.

24. Another example is a pile of salt. Does the pile contain a lot or a little salt? Let's say we open a small bag of salt obtained at a famous American fast-food restaurant and pour the contents on the table. There might be a gram or two of salt spilled out. One person looks at the pile (maybe a child) and exclaims "Look at all the salt!" Another, maybe an older sibling says "Oh, that's not such a big deal. That's a tiny pile." So, one child thinks there is a lot of salt, and the other doesn't. Their extensions for the concept of "how much salt" are not consistently determined.

25. Likewise, two concepts with the same extensions may not have the same intension. Consider two sets of integers defined as follows: A = {the greatest integer less than 5} and B = {the least integer greater than 3}. Both these sets define concepts, but their intensions, the essential characteristics, are different. The extensions, {4}, are the same.

26. Sometimes, the extension of a concept changes over time. This is why the definition for equality above contains the phrase "consistently determined". Requiring that the extension always have exactly the same objects is too strong a condition. The concept of "valid ZIP[4] codes" used by the US Postal Service is a case in point. ZIP codes change over time. Some new ones are added, and others are removed. At any point in time, two people should be able to determine all the current valid ZIP codes. However, if the comparison is done over a long enough time period, chances are good the lists will be different.

27. Now, it may be the case that for some concepts everyone agrees on the extension for all time. A simple example is the concept "even positive prime integers". Well, we get agreement as long as the concept (the intension) is understood. Not everyone remembers what a prime number is! But, agreement is certainly a subset of being consistently determined.

---

[3] This is deliberate. We want to avoid the use of the word "data" for now.
[4] ZIP stands for Zone Improvement Plan, instituted in the 1960's in the United States.

C. Partitions

28. Now, a concept with equality has a very important feature. It can used to partition a set of objects objectively (i.e., independent of who is doing the partitioning), where some of the objects are in the extension of the concept and others are not. A simple example is the concept "even positive numbers". The extension is the well-known set E = {2, 4, 6, …}. The concept partitions the positive integers into even and odd numbers. The reason the concept works as a means to partition the integers is that the extension is consistently determined.

29. Take a concept without equality, such as the nearness concept described above. The extension of the concept nearness, as applied to furniture, is all instances of pairs of pieces of furniture that are near. This extension undergoes changes almost continuously, however this is not the problem. As seen above, two people will not be able to consistently determine the extension even at the same point in time. We will not be able to partition objects objectively with such a concept.

D. Values

30. The concept of five, defined by "having five elements", is a concept with equality. Given a set of elements, it is easy to determine whether it has 5 elements or not. The same is true for any other number. The letter "A", for example, requires the concept "being the letter A". The same is true for other letters, symbols, words, and alphanumeric strings.

31. In every case, the concept partitions, and when that happens we refer to these cases as values. For example, five is "the value assigned to the number of elements in this set {a, b, c, d, e}." Note, the fact that five is a concept with equality now means the value five is useful. Each time five is used as a value, we are sure the extension of the concept, the instances of five, is determined in the same way - consistently.

32. Consider another example that might be a little more illustrative: The concept male, defined by being of the male sex. We apply this to people, and note that people are either male or female. That isn't quite the case, but for practical purposes it is. Male is a concept with equality, and just as in the example of five, we can use it as a value. Thus,

Definition: **Value** is a concept with equality

33. A value partitions the set of objects under study into 2 classes: Those objects that are in the extension of the value, and those that are not. Taking several values at once, it is clear that if the values are carefully chosen, one may construct a partition in which each class is a value. An example is the classification based sex with the values male and female.

34. Since values may be associated with numbers (e.g., integers, real), then the partition might have an arbitrarily high number of classes. Mathematically, a set of numbers may be infinite, but computationally, the set must be finite. Computers and people can't process an infinite number of things. Using a set of number values as the classes of a partition is not typical. For the purposes of this paper, though, we will do it.

35. Examples of partitions based on number values include the following two cases, with one for discrete data and the other for continuous:

- {1, 2, 3, 4, …125} is used to record the age of a person since his last birthday
- {all real numbers between 0 and 1 with a precision of 4 digits} is used to record probabilities - note, there are 10,000 such numbers, not an infinity

E. Datum

36. To summarize, a value is a concept with equality. Since, concepts are known and communicated through their definitions, then we need to define values, too. We refer to this definition as the value meaning.

37. Typically, we designate values many ways. For instance, the value male may be designated 'M', 'male', '0', etc. One of these designations is a term (male), and that is how we refer to the value. Confusingly, the term 'value' is also used to refer to the designations of values. E.g., 'M' as called a value even though it is really the designation of one.

38. A datum is how we refer to the designation of a value. It has the characteristic that it represents the class of some partition. Each class in the partition is a value, and the datum is a stand-in for the value. Since a value is a concept, it may have a designation. Therefore,

Definition: **Datum** is the designation of a value

39. In the modern theory of terminology, many concepts exhibit what is known as prototyping (Lakoff, 2002). The idea is that some objects in the extension of the concept are a better fit for the characteristics than others. As described above, a robin is viewed as a more prototypical bird than a penguin. That isn't to say a penguin is not a bird. It just doesn't fit the concept as well.

40. This effect can even be seen with values. Not all concepts used as values in practice quite fit the definition of a value given above. Data are used for many purposes, and compromises must be made to try to measure society. An example is the categories in an industrial classification system, such as the North American Industrial Classification System (NAICS). Domino's Pizza® is a chain of pizza making and delivery services throughout the US and other countries. Each shop is involved in 2 separate business activities: making and delivering pizza. The NAICS classification is based on production (what is done rather then inputs or outputs), so either cooking or delivering pizza is applicable in this case. Domino's Pizza® shops don't correspond well to either the pizza making industry or the pizza delivery industry. An arbitrary decision was reached on how to handle this (delivery), but this is the case of an object not fitting the scheme very well.

41. The example shows that values sometimes are subject to prototyping as well as concepts in general. Prototyping is an inherent attribute of concepts in language, and it arises when determining the extension of the concept. Since a value is a concept (with equality), then it may be subject to prototyping as well.

42. There are concepts that don't exhibit prototype effects. One is the concept "even positive prime integers." It is easy to see that the extension of this concept is well-defined; it is {2}.

43. The effect of this is that it is often hard to determine the correct value within a partition for objects, and this is inherent. This difficulty is called measurement error, and it is a problem in almost all scientific and statistical data collection activities. Though, it is possible to mitigate the effect by improving definitions and the collection process, as there are undoubtedly errors in collection itself, we also need to see that the problem is inherent to data.

F. Context

44. As we defined above, the concept associated with a datum is its value. There are also a number of other concepts associated with a datum that give it more meaning. These are

described in detail in Gillman and Johanis (2006) and Gillman (2006), however, from the perspective of a survey statistician, they are some of the following:

- Population
- Characteristic
- Category, class, or unit of measure
- Datatype
- Time, space, and survey

45. This list is not complete, but these concepts along with the value form a concept system. This concept system answers the questions "who", "what", "when", "where", "how", and "why" for the datum. We refer to these as the W5H questions. Therefore,

Definition: **Context** is the concept system associated with a datum that answers, minimally, the W5H questions

46. An example is the following:
4.7% is the unemployment rate for the US in January 2006. The details, corresponding to the bulleted items in the paragraph above are

- US labor force
- Unemployment
- Rate as a percent
- Real number
- January 2006, all of the US, and Current Population Survey

These answer the context questions (at least partially!) posed above.

47. We say the context is a concept system because it contains a set of concepts, which is evident, and it contains some relations between those concepts. For instance, the characteristic is related to the population, since it is a characteristic in the terminological sense of the concept represented by the population. The rate is a measure for representing the characteristic, unemployment; the rate has a real number datatype; and the value designated by the datum is a real number. This is a partial list, but the depth of the context is apparent.

G. Data Element

48. Since we have defined datum and context, now we can supply a definition of data element. Intuitively, a data element is a collection of datums that share similar semantics. Often, the notion that each datum cannot be further decomposed is added.

49. Consider the following example of a datum described as in the previous section:
A man in the US responds to the Current Population Survey in March 2006. He responds that he is a male to the questions about his sex. The details, corresponding to the bulleted items in section 3.2 are

- US adult person
- Sex
- Male
- Character
- March 2006, all of the US, and Current Population Survey

50. What does a data element containing this datum look like? As we stated above, it is a set of datums with shared semantics, i.e., sharing the same context. But if all the datums share the entire context, then they will all be the same value, i.e., all male.

51. A value is part of the context for a datum, however, it changes from datum to datum. All the rest of the context is shared. Now, a data element need not contain actual datums,

only that it could. This corresponds to a database table that is not yet populated, but the database designer knows what it means. As long as the shared context is known, then the data element can be said to exist. Therefore,

Definition: **Data element** is a set of datums that have the same context except the value[5]

H. Metadata

52. Metadata is often defined to be "data about data." This is not very precise. A better definition says metadata is data that are used to describe. This gets to the essential characteristic of metadata, they serve to describe something. The something is deliberately vague. Metadata can be used to describe anything. Here, we care about metadata describing data.

53. Our definition of datum leads us directly to a new definition for metadata. The context of a datum is the concept system. A concept system is a collection of concepts and the relations among them. The concept system associated with data can be reified, i.e. their meaning may be written out rather than kept as concepts in the heads of individuals. The reification is data, and this data describes the datum. It is metadata. Therefore,

Definition: **Metadata** for a datum are the data reified from the context of the datum

54. Now, there are obvious problems with this definition. It does not convey the essential characteristics of metadata. The definition, data that are used to describe, does this. The point is to convey a greater understanding of the definition of datum, and especially the role of the context. More will be said about this in the next section.

J[6]. Information

55. Information is a term that is widely and loosely used. Information is everywhere. The Internet is known as the Information Highway. All media are said to convey information to their audiences. So, what is information?

56. There are at least 2 well-known definitions of information, due to Langefors (1995) and Shannon (1948). Here we will look at the definition provided by Langefors. He says that information is an interpretation of data, and defines a function, the infological equation, to explain what interpretation means. The function is

$$I = i(D, S, T)$$

Where

- I is the information
- D is data
- S is pre-knowledge
- T is time
- i is the interpretation function

[5] In fact, a full description of a data element will include all the possible values and the signs designating them.
[6] The use of the letter I for a sub-section heading could not be distinguished from the Roman Numeral for the number one. So we skipped to the letter J.

57. In general, S is considered to be the total life-experience of the individual making the interpretation. An interesting consequence is that not everyone will interpret the same data in the same way at a given time.

58. The definition of datum given above, allows us to make a different definition of information, but one that is consistent with Langefors. First, our approach is limited to one datum at a time. Second, we don't define a process, the interpretation defined in the infological equation, but we know where the information is.

59. Each datum is a designation of a concept. Associated with that concept, the value, is a concept system, called the context. So,

Definition: The **information** represented by a datum is its context

V. CONCLUSION

60. The purpose of this paper was to provide and justify a new definition of data and some other concepts. We provided a precise definition of datum and derived definitions for context, data element, metadata, and information, as consequences. The relevant definitions are given here for easy reference:

- A concept has **equality** if its extension is determined consistently
- **Value** is a concept with equality
- **Datum** is the designation of a value
- **Context** is the concept system associated with a datum that answers, minimally, the W5H questions
- **Data element** is a set of datums that have the same context except the value
- **Metadata** for a datum are the data reified from the context of the datum
- The **information** represented by a datum is its context

61. The terminology theory is applied to define what a datum is: a designation of a concept with equality. And, each of the terms in the definition are defined as well. The concept piece for this is used to define context, data element, and information. Metadata is defined from context itself.

VI. REFERENCES

Bradley, W. (2004). *From Data to Knowledge: Standards and Systems for Sharing Data and Information Across Systems*. Working Paper #5 presented at the UNECE Workshop on Statistical Metadata,. Geneva, Switzerland.

CEN. (1995). *Medical Informatics - Categorical Structures of Systems of Concepts*. Draft. Brussels: European Committee for Standardization.

Date, C. (2003). *An Introduction to Database Systems (8th ed)*. Addison Wesley.

English, L. (1999). *Improving Data Warehouse and Business Information Quality*. New York: John Wiley and Sons.

Frankfurt, H. (2005). *On Bullshit*. Princeton: Princeton University Press

Froeschl, K., Grossmann, W., & Del Vecchio, V. (2003). *The Concept of Statistical Metadata*. Deliverable #5 for MetaNet Project. Retrieved July 2004 from http://www.epros.ed.ac.uk/metanet/deliverables/D5/IST-1999-29093_D5.doc.

Gillman, D. (2006) *Theory and Management of Data Semantics*. In D. Schwartz (ed.) Encyclopaedia of Knowledge Management. Hershey, PA, USA: Idea Group.

Gillman, D. and Johanis, P. (2006). *Metadata Standards and Their Support of Data Management Needs*. Working Paper #7 presented at the UNECE Workshop on Statistical Metadata,. Geneva, Switzerland

ISO (1993). *ISO/IEC 2382-1: Information technology - Vocabulary, Part 1: Fundamental terms*. Geneva: International Organization for Standardization and International Electrotechnical Commission.

ISO. (1999). *ISO 704: Principles and methods of terminology*. Geneva: International Organization for Standardization.

ISO. (2000). *ISO 1087-1: Terminology – Part 1: Vocabulary*. Geneva: International Organization for Standardization.

Lakoff, G. (2002). *Women, Fire, and Dangerous Things* (Reprint edition). University of Chicago Press.

Langefors, B. (1995). *Essays on Infology*. Stockholm: Studentlitteratur

Ogden, C. and Richard, I. (1989). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt.

Sager, J. (1990). A *Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

Wedberg, A. (1982). *A History of Philosophy - Vol 1: Antiquity and the Middle Ages*. Oxford: Clarendon Press