Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 267–273
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Conceptualizing the world: Lessons from history

Alexa T. McCray

Center for Clinical Computing, BIDMC, Harvard Medical School, Department of Medicine, 330 Brookline Avenue, Boston, MA 02215, USA

Received 21 June 2005
Available online 28 September 2005

Abstract

Throughout history, philosophers, scientists, and other scholars have named and organized the salient elements of the world. These efforts have led to conceptualizations that differ widely in both content and form. This paper argues that all conceptualizations are biased, both because they depend on the purposes for which they have been created, and because they are closely tied to the world view of their designers. This bias needs to be recognized, and its consequences need to be addressed if the conceptualizations are to be used for purposes other than those for which they were designed. The paper begins with a brief overview of the disciplines that have been concerned with conceptualizing particular domains. This is followed by a non-exhaustive, illustrative, historical perspective, and the paper concludes with the interesting case study of biological taxonomy.

Keywords: Metaphysics; Classification; Knowledge representation; Semantics; Terminology; Concept formation; History

1. Introduction

For over two millennia there have been efforts by philosophers, scientists, and other scholars to name and organize the salient elements of the world as they understood it. These efforts have led to conceptualizations that differ not only in their content and form, but also, and, importantly, in the purpose for which they were developed. Gruber has defined a conceptualization as "...the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them... A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose." [1, p. 907]. This definition has been criticized on the grounds that it is too vague, and that it has led to the circumstance that a wide range of constructs have become "common bedfellows" under a common rubric [2, p. vi]. Nonetheless, and perhaps because Gruber's definition allows a broad interpretation, it is the definition of conceptualization that is used here. The world that has been captured by these conceptualizations can be as broad as the universe itself, or it can be as narrow as a highly restricted domain, such as the world of restaurant dining [3].

It is the thesis of this paper that it is necessarily the case that *every* conceptualization is biased. This is because representing, or categorizing, the world depends on at least two crucial factors (1) the purpose for which the conceptualization is being created, and (2) the world view of its designer, with the corollary that this depends on the state of general knowledge at the time, as well as on the personal knowledge of the designer. There is bias in the choice of the categories themselves, in the hierarchical arrangement of the categories, in the depth or level of granularity chosen, and in the interrelationships that may or may not be made explicit among the categories. Bias is understood here to be a predisposed tendency toward a certain point of view, which is most often based on a particular system of beliefs. Bias does not immediately make a conceptualization incorrect or nonsensical, nor does it in any way need to diminish the usefulness of the conceptualization. The bias simply needs to be recognized, and its consequences need to be addressed if the conceptualization is to be used for some purpose other than the one for which it was originally designed.

E-mail address: alexa_mccray@hms.harvard.edu.

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.08.007

The paper begins with a brief overview of the disciplines that have been concerned with conceptualizing, or categorizing, particular domains, these disciplines being closely allied with the purposes for which the conceptualizations have been designed and the types of artifacts that result from these efforts. This is followed by a non-exhaustive, illustrative, historical perspective. The paper concludes with the interesting case study of biological taxonomy, which illustrates many of the issues raised in this paper.

1.1. Conceptualizing the world

It is fair to say that all attempts at categorizing the world have as their goal an accurate representation of some reality, in some domain of interest. Philosophers, for example, have and continue to be concerned with fundamental notions of existence. Philosophical ontology, according to one view, is "the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality." [2, p. iii]. The goal may be to understand the existence of a supreme being, to understand the nature of existence, or, more broadly, to understand the nature of reality. Philosophers since the time of Aristotle have grappled with the relationship between language, thought, and reality, and, in modern times, the views of Frege [4], de Saussure [5], Ogden and Richards [6], and Wittgenstein [7] have been particularly influential. Ogden and Richards, for example, building on the work of both Frege and de Saussure, propose the "semiotic triangle," such that there is a referent (in reality), a symbol (language) that may refer to that referent, and thought (the concept as it exists in the brain). Characterizing how conceptual knowledge is organized in the brain continues to be of interest to both linguists and psychologists. Linguists such as Dornseiff [8] and Trier [9] propose that concepts are grouped together in the brain according to certain shared characteristics, and psychologists in their attempts to understand the nature of cognition, have posited anatomical loci for semantic fields, or domains, in the brain [10]. A New York Times article in the early 1990s reported, "For a century or so, psychologists pondering the brain's memory handling system have suspected that the brain had some system of putting information in categories, with a separate pigeon-hole for categories like dogs, plants, or numbers, each in a separate network of cells...in the past decade a new surge of information has led to the identification of 15–20 different categories that appear to be the brain's own natural categories for knowledge." [11]. Some neuropsychologists have posited a high-level distinction between the semantic domains of natural kinds, such as animals and plants, on the one hand and man-made objects, such as tools and vehicles on the other [12]. Lesions, disease processes and imaging techniques, such as PET scans and functional MRI's are all used as evidence [13]. Recent findings, however, seem to support an attribute-based neural organization of semantic knowledge, rather than basic categories [14].

Closely allied with work in psychology, computer scientists have, over the last forty years or so, attempted to model the world, or some particular portion of that world In some cases, the purpose has been to mimic human cognitive abilities, as in, for example, the CYC project which attempts to capture human common sense reasoning [15]. Often, however, the goal has been to build practical computational systems for applications such as machine translation, information extraction, question answering, and text summarization [16]. The resultant artifacts have been variously called frames, semantic networks, conceptual graphs, and, most recently, ontologies [17–20]. For Gruber an ontology is an "explicit specification of a conceptualization" [1, p. 907], and, similarly, for Poli it is a framework or structure "within which catalogues, taxonomies, terminologies may be given suitable organization." [20, p. 313].

Occasionally, particularly when researchers are working within similar or the same domain there is strong disagreement about what the correct representations should be. Nowhere is this more true than in standardized coding systems. Medical coding systems, for example, have been developed for epidemiological and data comparability purposes, e.g., the International Classification of Diseases of the World Health Organization [21], for billing purposes, e.g., the Current Procedural Terminology of the American Medical Association [22], and yet others have been developed for information indexing and retrieval, most notably the Medical Subject Headings of the National Library of Medicine [23]. Chute alludes to some of the disagreements over the years in the field of medical coding systems: "Many authors have derided one modality or another, failing to recognize that each serves its purpose." [24, p. 301]. Coding systems have in common that they regulate the terminology that is allowed [25]. The NISO thesaurus standard states, for example, that controlled vocabularies for information retrieval must be arranged in a specific order such that "...relationships among terms are clearly displayed and identified... Its purposes are to promote consistency in the indexing of documents, predominantly for post-coordinated information storage and retrieval systems, and to facilitate searching by linking entry terms with descriptors." [26, p. 38]. The standard not only has naming rules, but it also gives guidelines for the hierarchical arrangement of the terms and for allowable inter-term relationships. In spite of the prescriptive nature of these coding systems, there is rarely agreement on the resulting artifacts. There is often dissatisfaction with the concepts chosen, the granularity with which the concepts are represented, and the hierarchical or relational structure into which the concepts have been placed.

Work on biological taxonomies shares some of the same problems. Biologists, in the pursuit of knowledge about living beings, have developed taxonomies to represent biological knowledge as they understand it. Arguments about how to accomplish this have been in evidence since even before the time of Linnaeus. Ereshefsky notes that one area of disagreement is whether a biologic entity should be rep-