Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-68

An Information Artifact Ontology Perspective on Data Collections and Associated Representational Artifacts

Werner CEUSTERS[a,1]
[a]*New York State Center of Excellence in Bioinformatics & Life Sciences, Buffalo, USA*

Abstract. Biomedical data collections are typically compiled on the basis of assessment instruments and associated terminologies and their data structure explained by means of data dictionaries. The Information Artifact Ontology (IAO) is an attempt to give a realism-based account of the essence of information entities and how components of such entities relate to each other and to that what they are information about. Changes in the taxonomy and the definitions of the IAO, most importantly the addition of the terms 'representational artifact' and 'representational unit', are proposed to make the IAO a useful tool to clarify formally the distinctions and commonalities between data collections and associated artifacts that are compiled independently from each other, yet cover the same domain.

Keywords. Ontological Realism, Information Artifact Ontology, Terminology, Representations

Introduction

The goal of the OPMQoL-project funded by the National Institutes of Health (NIH) is to obtain better insight into the complexity of pain disorders, specifically concerning the assessment of different pain types in the orofacial region, as well as into pain-related disablement and its association with mental health and quality of life.

Five existing data collections compiled independently from each other in respectively the US, Germany, Sweden, the UK and Israël, and covering in total 2000 patients, have been made available for this study. The data collections cover the same domain, but are distinct in various respects: (1) some variables are identical across collections, others involving, for instance, somatization, depression and anxiety, are different because measured with in total 22 distinct assessment instruments; (2) these instruments contain each between 50 and 500 unique assessment items, but, although frequently sharing intent, do not share a similar presentation across forms, supporting detail, instructions regarding the sources of information that can be used to complete each item, or severity/frequency response scales that are comparable across instruments; (3) because of their distinct origins, the data collections incorporate cultural influences related to pain report that have an impact on the comparability of the collections, despite the use of common instruments.

[1] Corresponding Author. Werner Ceusters. Ontology Research Group, New York State Center of Excellence in Bioinformatics & Life Sciences, and Department of Psychiatry, University at Buffalo, 701 Ellicott street, Buffalo NY 14204, USA; E-mail: ceusters@buffalo.edu.

One specific aim of the project is to make these data collections comparable by building a realism-based reference ontology for pain-related disablement, mental health and quality of life (OPMQoL) following the principles of Ontological Realism [1], a methodology for the coordinated evolution of biomedical ontologies which is embodied in the Basic Formal Ontology and used in over hundred projects and institutions world-wide [2].

The work reported on here consisted of the first step in this endeavor its purpose being to obtain a clear understanding of how the various information sources made available to the project relate to each other, and how that understanding can contribute to further advancing our insight in how information in general precisely relates to that what it is information about. The challenge here is thus to align the terminological perspective according to which the assessment instruments and data collections are designed on the one hand with the ontological perspective on the other hand, and this, in addition, in line with the principles of Ontological Realism. There are currently two efforts that embrace Ontological Realism in their attempt to get a better grasp on what representational artifacts such as terminologies, ontologies, and data collections exactly are. One is a terminological effort initiated by Gunnar Klein, former chairman of CEN TC 251, which delineates the boundaries between concept systems and ontologies and which holds some promises towards harmonization without however any clear indication on how such harmonization could be achieved [3]. The other one, the Information Artifact Ontology (IAO), is an ontological effort to describe the distinctions and commonalities between various sorts of information entities [4].

1. Methods

The available data collections, their data dictionaries and some of the assessment instruments, corresponding terminologies and coding manuals - all together from here on called 'the sources' - used for these collections were therefore analyzed in function of the IAO and Gunnar Klein's proposal, thereby further taking into account earlier work on the nature of representational units (RUs) and what sorts of entities such units might stand for [5-6]. The most generic types of compositional elements of the sources and the sources as a whole themselves were then defined and classified in the taxonomy of the IAO and the relationships amongst them further clarified in a UML-diagram. Where deemed required, RUs were added to the IAO and modifications to existing IAO definitions proposed.

2. Results

Table 1 shows a proposal for an extended IAO taxonomy ('Term'-column) with corresponding definitions ('Definitions'-column), thereby incorporating most of the types of elements instances of which are the building blocks of the sources. Terms in the 'Term'-column depicted in **bold** are additions to the original taxonomy, with the exception of ***Term*** which IAO thus far underspecified as 'part of an ontology'. It is for each definition indicated whether (1) it is taken verbatim - modulo minor changes that do not change the intended meaning - from a referenced source, (2) adapted from a source, this adaptation being such that it follows the principles of Aristotelian definitions, or (3) newly introduced (in case no reference is provided).

Table 1: Proposal for an extended IAO taxonomy and corresponding definitions

Term	Definition
Information Content Entity (ICE)	an *entity* that is *generically dependent* on some artifact and stands in relation of *aboutness* to some *portion of reality* [4]
Representational Artifact (RA)	an **ICE** which is believed to *represent* a *portion of reality* external to the representation (modified from [5])
Representational Unit (RU)	a **RA** which according to the structural conventions it is designed, is not built out of any other **RAs**
Denotator	a **RU** which *denotes* directly an *entity* without providing a description [6]
Term	a **RU** which is a general expression in some natural language used to refer to *portions of reality* (modified from [5])
Composite Representation	a **RA** built out of constituent sub-representations as its parts (modified from [5])
Data Collection	a **composite representation** built out of **measurement data**
Data Dictionary	a **composite representation** describing, inter alia, what **data items** in a **data collection** are *about*, including a **data format specification**
Terminology	a **RA** consisting of **terms** (modified from [5])
Ontology	a **RA** comprising a taxonomy as proper part, whose **RUs** are intended to designate some combination of *universals*, *defined classes*, and certain *relations* between them [3]
Realism-based Ontology	an **ontology** built out of **RUs** which are intended to be exclusively about *universals* and certain *relations* between them, intended to mimic the structure of reality, and which correspond to that part of the content of a scientific theory that is captured by its constituent general terms and their interrelations [3]
Reference Ontology	an **ontology** intended to provide an *informationally complete* representation of a domain
Application Ontology	an **ontology** representing the *portion of reality* which is relevant for some purpose in some community
Assessment Instrument Ontology	an **application ontology** describing the *portion of reality* covered by an **assessment instrument**
Data Collection Ontology	an **application ontology** describing the *portion of reality* covered in a **data collection**
Data Item	a **RA** that is intended to be a truthful statement about something (modulo, e.g., measurement precision or other systematic errors) and is constructed/acquired by a method which reliably tends to produce (approximately) truthful statements (modified from [4])
Measurement Datum	a **data item** that is a recording of the output of a measurement. [4]
Directive Information Entity	an **ICE** whose *concretizations* indicate to their *bearer* how to *realize* them in a *process* [4]
Conditional Specification	a **directive information entity** that specifies what should happen if a trigger condition is fulfilled [4]
Rule	an executable **conditional specification** which guides, defines, or restricts actions [4]
Bridging Axiom	a **rule** specifying how a **RA** should be interpreted in terms of an **application ontology**
Data Format Specification	the information content borne by the *document* published defining the specification (modified from [4])
Plan Specification	a **directive information entity** that when *concretized* is *realized* in a *process* in which the *bearer* tries to achieve the objectives, in part by taking the actions specified [4]
Assessment Instrument	a **plan specification** designed to compile **data collections** reliably, validly and reproducibly

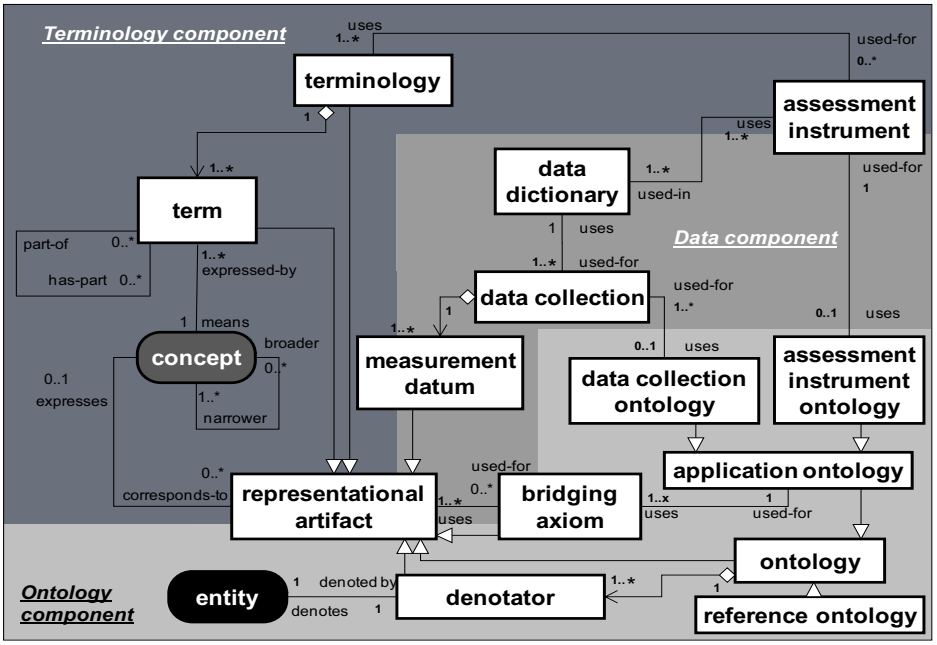

Figure 1: relationships amongst sources and their components.

Terms in **bold** in these definitions are defined elsewhere in the table, whereas terms in *italic* are additional technical terms outside the realm of information artifacts for which all explanations cannot be provided here because of space limitations but can be found elsewhere [1, 7]. Essential for the understanding of the proposed definitions and the relationships depicted in Figure 1, are nevertheless (1) *concept*: meaning of a term as agreed upon by a group of responsible persons [3], (2) *entity*: anything which is either a universal or an instance of a universal [3], and (3) *portion of reality*: any entity or configuration of entities standing in some relation to each other [6].

Additional relationships amongst the types of elements defined in Table 1 are depicted in Figure 1 which follows standard UML conventions for the relations, all of which have specified cardinalities: solid-arrowed lines stand for subsumption, the arrow pointing towards the subsumer; arrows with squares stand for composition, the arrow pointing towards the component; and un-arrowed lines representing associations which are named in both directions, the name printed close to the range of the relation.

3. Discussion and conclusion

The core elements in the proposal advanced here, and missing in the IAO, are *Representational Unit* (RU) and *Representational Artifact* (RA). The motivation to include RA as a direct subsumer of *Information Content Entity* (ICE) is the distinction between 'just' *being about* a portion of reality and *representing* a portion of reality. False or misleading information is still *about* something, but does not *represent* that something. This addition, combined with replacing '… *about something*' in the original definition with '… *about a portion of reality*', would also avoid the misunderstanding expressed in [8] that *aboutness* would tie an ICE to an *entity*. And it would also allow

the various types of sources and data collections to have an appropriate place in the taxonomy without harmful underspecification. The proposal does however not accommodate those who perceive fictional stories as ICE too since fictions aren't about anything at all.

The addition of RU in the IAO would offer a possibility to bridge the gap between terminologies and concept systems on the one hand and ontologies on the other hand. Although [3] gives a clear account of what this gap exactly is and why it should be maintained, it does not offer a solution for applications that have to integrate/interface instances of both of these types of resources while still embracing Ontological Realism. Because, as proposed here, both *terms* (used in terminologies, assessment instruments and data dictionaries) and *denotators* (denoting particulars when components of a data collection, or universals when components of ontologies) are RUs, they can both be used in *bridging axioms* that formally describe how *data items* clarified in terms of a terminology can be translated into a representation that exclusively uses *denotators*, and this without resorting to description language dialects that are inconsistent with Ontological Realism [9].

Acknowledgements

The work described is funded in part by grant 1R01DE021917-01A1 from the National Institute of Dental and Craniofacial Research (NIDCR). The content of this paper is solely the responsibility of the author and does not necessarily represent the official views of the NIDCR or the National Institutes of Health. Thanks also to Alan Ruttenberg, custodian of the IAO, for useful comments.

References

[1] Smith B, Ceusters W. Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies. Applied Ontology. 2010; 5(3-4):139-88.
[2] Stenzhorn H. Basic Formal Ontology (BFO) users. 2011 [cited 2012 January 26]; Available from: http://www.ifomis.org/bfo/users.
[3] Klein GO, Smith B. Concept Systems and Ontologies: Recommendations for Basic Terminology. Japanese translation in Journal of the Japanese Society for Artifical Intelligence 2010; 25:433-41.
[4] Information Artifact Ontology. 2012 [cited 2012 January 24]; Available from: http://code.google.com/p/information-artifact-ontology/.
[5] Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. KR-MED 2006, Biomedical Ontology in Action. Baltimore MD, USA 2006.
[6] Ceusters W, Manzoor S. How to track Absolutely Everything? In: Obrst L, Janssen T, Ceusters W, editors. Ontologies and Semantic Technologies for the Intelligence Community Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press; 2010. p. 13-36.
[7] Smith B. Basic Formal Ontology 2.0. 2012 [cited 2012 January 24]; Available from: http://ontology.buffalo.edu/bfo/Reference/.
[8] Hastings J, Batchelor C, Neuhaus F, Steinbeck C. What's in an 'is about' Link? Chemical Diagrams and the Information Artifact Ontology. In: Smith B, editor. Proceedings of the International Conference on Biomedical Ontologies. Buffalo NY, 2011. p. 201-8.
[9] Schulz S, Brochhausen M, Hoehndorf R. Higgs Bosons, Mars Missions, and Unicorn Delusions: How to Deal with Terms of Dubious Reference in Scientific Ontologies. In: Smith B, editor. Proceedings of the International Conference on Biomedical Ontologies. Buffalo NY, 2011. p. 183-9.