Grant R01DE021917 from the National Institute of Dental and Craniofacial Research
Project Period: 07/01/2011 – 06/30/2014

An Ontology for Pain and related disability, Mental health and Quality of Life (OPMQoL)

Final Report

September 27, 2014

Werner Ceusters

University at Buffalo, Department of Biomedical Informatics
Ontology Research Group (ORG), New York State Center of Excellence in Bioinformatics & Life Sciences
UB Institute for Healthcare Informatics
923 Main street
Buffalo NY 14203
USA

ceusters@buffalo.edu
office: (716) 881-8971
mobile: (716) 418-4237

Table of Contents

1 Executive Summary

1.1 Vision and research goals

The broad, long-term vision underlying our research over the past ten years is one in which representational artifacts designed for use in software applications mimic the structure of reality to the best understanding of their authors. This holds for artifacts that represent generic information such as classification systems, terminologies and ontologies as well as for data repositories such as electronic health records, clinical research datasets and data warehouses. And it holds not only for what is believed to be the case today, but also for how matters have been in the past.

The research carried out under this grant aimed to advance the state of the art in representing the complexity of pain disorders, specifically concerning the assessment of different pain types as well as pain-related disablement and its association with mental health and quality of life. The goal is to develop an ontology which is then used to integrate available clinical research datasets that broadly encompass the major types of pain (orofacial pains, temporomandibular disorder pain, and headache) recognized to occur in the oral and associated regions and incorporating a broad array of measures consistent with a bio-psychosocial perspective regarding pain. The datasets cover the same domain, but are distinct in various respects: (1) some variables are identical across datasets, others involving, for instance, somatization, depression and anxiety, are different because measured with distinct instruments; (2) the data exhibit different levels of granularity; (3) because of their distinct origins (US, UK, Sweden, Israel, and Germany), the datasets incorporate cultural influences related to pain report that have an impact on the comparability of the data sets, despite the use of common instruments.

Our hypothesis is that the ontology will make it possible to describe the datasets in a uniform and formal way, and be general enough to include other datasets in the same domain once they become available. The importance of this endeavor lays in its contribution to solving an important problem, namely that the phenotype of many pain conditions is insufficiently defined in terms of the scope, the natural history and/or clinical course of the disease subgroup of interest, and, most importantly, with respect to disease traits for which laboratory research has provided important pathogenetic insight.

1.2 Progress towards originally stated aims

Our plan was to test our hypothesis through achievement of the following specific aims:

- Aim 1: describe the portions of reality covered by the five datasets and acquire broad consensus in the field with respect to its face validity.
- Aim 2: design the bridging axioms required to express the data dictionaries of the datasets in terms of the OPMQoL ontology and translate these axioms in the query languages used by the underlying databases.
- Aim 3: validate the ontology by querying the datasets with and without using the ontology and by comparing the results in function of the clinical question identified.
- Aim 4: document the development and validation approach in a way that other groups can re-use and expand OPMQoL, and use our approach in other domains.

The following summarizes our results in terms of the tasks as originally conceived in the proposal to achieve these aims. Details about these results are provided in additional sections of this report and papers published as a result of this grant.

1.2.1 Data collection and preparation

- we obtained all datasets and supporting documents over a timespan from May 2011 to June 2014 (see p22),
- for one dataset, IRB issues had to be resolved; for some datasets data dictionaries were not available and had to be constructed; for one dataset, which existed only on paper, a data entry tool had to be developed (see p29),
- we trained a PhD student into realism-based ontology design and a Master student in ontology-based evaluation of diagnostic classifications for headache,
- four datasets required a considerable collaborative effort to understand the origin and meaning of the encoded variables, an effort that was underestimated in our work plan,
- a method was developed to check consistency of data values with the corresponding data dictionary (see p100).

1.2.2 Ontology and terminology development

1.2.2.1 Terminology alignment and mapping.

- We used ontological principles to demonstrate inconsistencies in the ways diagnostic classifications – in our case chapter XIII of the International Classification of Headache Disorders – are build today and proposed guidelines for improvement (see p47).
- An investigation into the ontological status of biomarkers was carried out. The current terminology was found to be not precise enough and recommendations for ontological definitions for biomarkers for orofacial pain were formulated (see p68).
- An experiment was carried out to assess to what extent an ontological reformulation of pain assessment terms could reduce overlap (see p60).
- The method developed to annotate assessment instruments and datasets allowed us to combine the originally separately proposed tasks of terminology alignment and mapping (see p99 and p38).

1.2.2.2 Development of application ontologies.

- We evaluated the Information Artifact Ontology (IAO) for its potential to represent the relationships between information sources of various types such as data collections, data dictionaries, assessment instruments, etc., in a coherent fashion and found that it is possible to do so on the condition that the IAO is modified along the lines suggested by our research (see p30 and p108).
- We analyzed several of the assessment instruments used in the datasets and supporting documentation provided to us and linked data types found therein to the reference ontology (see p38).

1.2.2.3 Development of the reference ontology OPMQoL

- Existing resources in the National Center for Biomedical Ontology BioPortal were tested for their suitability to serve partly or in total as reference for OPMQoL. Coverage of pain assessment terms was found to be insufficient and coherent definitions lacking, thus not an option (see p50),
- We participated in parallel work to establish an ontology for mental functioning and unexplained syndromes (see p34), areas of high relevance for orofacial pain syndromes,
- We combined several realism-based ontologies to form the upper domain ontology of OPMQoL (see p74),

- We linked data types found in the assessment instruments (see p46) and in the analyzed datasets (see p101) to types in the upper domain ontology.

1.2.2.4 Ontology and Terminology publishing

- All materials are implemented as Excel files (see examples on p39, p75 and p99).and available for release after validation and, for datasets, on the basis of data use agreements, perhaps subject to additional IRB approval, with the original data sources.

1.2.3 Data sets/ontology bridging

- We developed a mechanism to translate clinical research datasets – exemplified by the datasets used in this project – into self-explanatory datasets for easier integration (see p91) and filed this as an invention.

1.2.4 Validation

- Validation revealed that data types from sources are usually linked too high up in the upper domain hierarchy. Correction is ongoing.

1.2.5 Documentation

- Documentation has been produced for all tools developed (see Appendix).

1.2.6 Dissemination & awareness

- See list of papers and presentations in this section.

1.3 List of Publications and Presentations

1.3.1 Published papers

1. Hastings J, Ceusters W, Smith B, Mulligan K. Dispositions and processes in the Emotion Ontology. In: Bodenreider O, Martone ME, Ruttenberg A (eds.), Proceedings of the 2nd International Conference on Biomedical Ontology (ICBO-2011), Buffalo, NY, USA, July, 26-30, 2011:71-78. CEUR Workshop Proceedings, ISSN 1613-0073, available online at CEUR-WS.org/Vol-833/.
2. Hastings J, Ceusters W, Smith B, Mulligan K. The Emotion Ontology: enabling interdisciplinary research in the affective sciences. In: Beigl M, Christiansen H, Roth-Berghofer TR, Kofod-Petersen A, Coventry KR, Schmidtke HR (Eds.) Modeling and Using Context; Proceedings of CONTEXT 2011, Karlsruhe, Germany, September 26-30, 2011, Lecture Notes in Artificial Intelligence 6967;119-123.
3. Nixdorf D, Drangsholt M, Ettlin D, Gaul C, de Leeuw R, Svensson P, Zakrzewska J, DeLaat A, Ceusters W. Classifying orofacial pains: a new proposal of taxonomy based on ontology. Journal of Oral Rehabilitation 2012;39(3):161-169 (PMC3383028).
4. Ceusters W. An Information Artifact Ontology Perspective on Data Collections and Associated Representational Artifacts. Medical Informatics Europe Conference (MIE 2012), Pisa, Italy, August 26-29, 2012, Stud Health Technol Inform. 2012;180:68-72.
5. Doing-Harris K, Meystre SM, Samore M, Ceusters W. Applying Ontological Realism to Medically Unexplained Syndromes. 14th World Congress on Medical and Health Informatics (MEDINFO 2013), Stud Health Technol Inform. 2013;192:97-101. (PMID: 23920523 [PubMed - in process])

6. Selja Seppälä, Barry Smith and Werner Ceusters, “Applying the Realism-Based Ontology-Versioning Method for Tracking Changes in the Basic Formal Ontology”, Formal Ontology in Information Systems. Proceedings of the Sixth International Conference (FOIS 2014), Amsterdam: IOS Press, 227-240.

1.3.2 Forthcoming papers

1. Ceusters W, Hsu CY, Smith B. Generating Self-Explanatory Data Repositories from Clinical Research Datasets using Referent Tracking. International Conference on Biomedical Ontologies, ICBO 2014, Houston, Texas, Oct 6-9, 2014. (accepted)
2. Ceusters W. Pain Assessment Terminology in the NCBO BioPortal: Evaluation and Recommendations. International Conference on Biomedical Ontologies, ICBO 2014, Houston, Texas, Oct 6-9, 2014. (accepted)
3. Ceusters W. An alternative terminology for pain assessment. In Workshop on Definitions in Ontology, International Conference on Biomedical Ontologies, ICBO 2014, Houston, Texas, Oct 7, 2014. (accepted)

1.3.3 In preparation

1. Nasri-Heir C, Alnaas D, Eliav E, Brian E. Cairns B, Ceusters W. Biomarkers of Chronic Orofacial Pain: from research to clinic.
2. Ceusters W, Benoliel R, Durham J, Raphael KG, Michelotti A, Ohrbach R. Perspectives on Next Steps in Classification of Orofacial Pain – Part 1: Role of Ontology.
3. Raphael KG, Durham J, Benoliel R, Ceusters W, Michelotti A, Ohrbach R: Perspectives on Next Steps in Classification of Orofacial Pain: Role of psychosocial factors (Part 2)

1.3.4 Presentations

1. *Ontological Realism for Biomedical Ontologies and Electronic Health Records*. Tutorial as part of the Medical Informatics Europe Conference, MIE 2011, Oslo, Norway, August 28-31, 2011.
2. How to Overcome the Lack of Data Interoperability and Data Quality. Lecture as part of Utilization of Electronic Health Record (EHR) Data for Clinical Research, short pre-conference course for the 3rd Annual Summit for Clinical Trials Operations Executives (SCOPE), Miami, FL, February 6, 2012.
3. *Ontologies for the bio-science industry: development and use*. Short course at the Molecular Medicine Tri-Con 2012 conference, San Francisco, CA, February 20, 2012.
4. *Pain and Mental Health: a Case-Study in Information Driven Research*. Lecture as part of the Core Curriculum in Clinical and Translational Research Seminar Series, Buffalo, NY, February 22, 2012.
5. *Realism-Based Ontology for Integrating Individually Compiled Biomedical Data Repositories*. 3-Hour tutorial given as (1) part of the European collaborators meeting in Milan, Italy, Sept 1, 2012, and (2) part of the Medical Informatics Europe Conference (MIE 2012), Pisa, Italy, August 26, 2012.
6. *Referent Tracking: focus on particulars*. Lecture for PHI 531 Problems in Ontology, class 23893, University at Buffalo, Sept. 10, 2012.
7. *Ontology and Data Abstraction.* Lecture for the UB Advanced Graduate Certificate Program in Medical/Health Informatics *Introduction to Medical Informatics (Part 1)* – Fall 2012 – MHI501, University at Buffalo, Nov. 14, 2012.

8. *Ontology: innovative approach to orofacial pain classification*. IADR Satellite Symposium on Orofacial Pain Assessment: Classification, Biobehavior, QST, and Biomarkers, March 19, 2013, Seattle, WA.
9. *Biomedical Ontology and Referent Tracking: Introduction to Basic Principles*. IADR Satellite Workshop on Orofacial Pain, March 20, 2013, Seattle, WA.
10. *The principles of high quality ontology design*. Applied Mathematics Seminar, University at Buffalo, April 15, 2014.
11. *Ontology, TMD and beyond; Principles for Taxonomy Development*. ACTTION-APS Pain Taxonomy Meeting, Westin Annapolis, Annapolis, MD, July-18-19, 2014.
12. *Data Dictionaries for Pain and Chronic Conditions Ontology*. Investigators Meeting on Chronic Overlapping Pain Conditions, NIH Main Campus Bldg. 31, Bethesda, MD, September 16-17th, 2014.

1.4 Conclusion and Recommendations

The research conducted has provided us considerable insight in not only the power that realism-based ontology approaches have to offer in the analysis of complex domains such as pain and the way findings are represented in clinical research datasets, but also in the complexity of the approach itself.

The latter has indeed been argued to be a problem for the development of practical ontologies [1] but the arguments provided in that critique were easily refuted through analysis of the foundations for these arguments themselves and by the discovery of flaws in the ontology developed by means of an alternative approach [2]. The latter is also confirmed by the low quality we discovered as part of the research conducted here with respect to 'ontologies' submitted to the NCBO BioPortal (see p50).

However, we must confess that – so to say 'blinded by our expertise' – we underestimated the complexity of the approach. This became not only apparent while providing education to the master and PhD students that assisted us in this research, but also while communicating with collaborators in the project. There is an incredibly large difference between, on the one hand, the way pain research experts conceptualize the first-order reality they are dealing with, the constructs they develop to understand and explain their findings, and how they formulate theories and communicate about it through diagnostic criteria, classification systems and scientific publications, and, on the other hand, the way realist ontologists look at matters. Although it sometimes has been phrased to be a 'disconnect', it clearly is not: in every single case we have been able, with great motivation, effort and patience from either side though, to find the tiny bridge that connects the two worlds. This is for instance witnessed by the lengthy editorial processes that led to [3] and are still going on for the forthcoming publications mentioned above. The 'track changes' and multi-author features offered by modern text-processors will prove to be extremely valuable for a future publication on the challenges that researchers on either side of the bridge found themselves confronted with when not just going for the easy, reductionist solution typical for a consensus/common denominator type of work, but rather for one where everything that has to be said is indeed said, without leaving any room for ambiguities.

Reducing ambiguities in clinical research datasets was one of the main drivers of the research conducted here. Did we succeed? Partly for sure. Thinking in terms of what ontological realism dictates allowed us to discover a great deal of ambiguities in the analyzed assessment instruments and datasets, specifically with respect to the kind of particulars, i.e. individual entities, that exist on the side of the patient when an answer to a question from an assessment instrument is given in one or other way, or when a specific value is provided for a variable in a

dataset. When, for instance, one question is about the 'pain of longest duration' over the past six months, and another one about the most intense pain during that period, the problem for the realist ontologist is whether the patient, when providing the requested documentation, is referring to the very same instance of pain, or to two distinct ones: both type of situations may indeed occur and instances thereof will be different from one patient to another. Unfortunately, although this ambiguity can be detected, it cannot be resolved post-hoc. Does it matter? Pain experts might say it doesn't, but how can we know for sure if it has never been researched? Statistics alone will not give the right answers [4, 5].

A conclusion we can reach at this point is in any case that for future research, such ambiguities should be avoided. For new clinical trials, not only in the domain of pain research, a realist ontology based approach should be used to scrutinize proposed variables and constructs in order to free them from any ambiguities. Furthermore, data dictionaries should be developed ***prior*** to the data collection and subjected to the same type of ontological analysis. Perhaps a few prospective studies can be done where both approaches – the traditional one and the realist ontology-based one – are applied to matched cohorts. And for sure, authors of classification systems with or without diagnostic criteria, should seriously consider to apply the methodology to avoid the sort of problems encountered in the ICHD (see p47).

Given the enormous amount of effort required to provide a post-hoc ontological interpretation of existing research datasets – an effort of which we underestimated the size – the question remains whether it is worthwhile. We don't have the answer yet, but will continue to pursue it. We have therefor the support of the UB Institute for Healthcare Informatics which decided to build out its data repository along the lines described. We will further work on the materials developed here through projects with graduate students from the newly formed UB Department of Biomedical Informatics, and more specifically its Division of Biomedical Ontology whose task it is, amongst others, to develop better educational resources for realism-based clinical research.

2 Introduction

The work described in this report is a logical continuation of the research initiated by the PI in the early nineties, which aims:

(1) to bring unconstrained natural language understanding up to a level that it can be used for man-machine communication and

(2) to design software that is able to make data semantically interoperable for automated decision support.

This research has primarily been focused around methods and techniques for overcoming the burdens associated with traditional paradigms for structured documentation in electronic patient records [6-11]. Central to our earlier work is the vision that, to understand natural language and structured patient data, software programs must incorporate knowledge about how the world is structured, how this structure is perceived by humans, and how humans communicate about it [12-14].

We found that ontologies, primarily those based on sound philosophical theories, are essential components for providing this sort of knowledge, and in such a way as to do justice to the difference concerning what is the case and what is known or believed to be the case [15].

The word 'ontology' is used for various types of artifacts created and used in different communities to represent those entities and relationships salient to a given domain. Such artifacts range from formal upper-level ontologies expressed in first order logic to the simple user-defined keyword lists used, for example, to annotate resources on the Web. In between are taxonomies and controlled vocabularies such as MeSH, often used for information indexing and retrieval, and whose organization is primarily hierarchical, as well as ontologies and vocabularies which represent also non-hierarchical relationships such as the Foundational Model of Anatomy [16-18], SNOMED-CT [19-22] and the NCI Thesaurus [23-28].

Increasingly, ontologies are being used to support the retrieval, integration and analysis of a variety of different kinds of biomedical data. Ontology-based technology has been successful especially in support of data-driven research in the basic biological sciences and in model organism studies, and efforts are now being made to extend these successes to the domain of human disease and diagnosis. The most successful ontologies, above all the Gene Ontology [29], rest on objective classifications of biological phenomena primarily at the molecular and cellular levels. In other areas, however, we face difficulties in applying the same approach specifically where we are dealing with clinical data pertaining to pain and other symptoms of human disease that are marked by the feature of subjectivity.

2.1 Vision

We embrace the vision that every assertion which is concretized in an information system and which is of value to contribute positively to an individual's well-being in particular or to the advance of biomedical science in general should be instantly available and usable in any other information system – wherever located – that is capable to capitalize on that value, thereby respecting all privacy, security, legal and moral restrictions that are applicable to it, not only with respect to the persons or organizations about which the assertion is made, but also to those that generated it or had access to it.

Mainstream approaches that aim to achieve the goals of this vision include data exchange through purpose-specific messages, regimentation of data types and data collections through common data elements, and concept-based terminologies and ontologies to reduce the lack of semantic interoperability and provide mappings – usually incomplete – between knowledge and data sources. These approaches, though contributing to some extent to the overall endeavor,

are insufficient because of the overemphasis on the information structures through which data and knowledge are gathered, made accessible and processed without aligning it with the structure of reality the data and metadata are about. They also do not take into account various implicit elements of information, for instance whether data that seem to be missing from a repository are justifiably absent (e.g. certain questions are not asked depending on the answer to previous questions) or truly missing (e.g. answers to questions that were asked but for some reason not being recorded). The consequence is that certain biomedical research questions cannot be answered.

Key research questions to be addressed through ontology research include:

1. How large a portion of existing big data and knowledge collections used in biomedicine can, using existing terminological, ontological and metadata standards, be automatically transformed into self-explanatory information repositories that describe unambiguously (1) what the data they contain are about, (2) to what extend the data are faithful to the corresponding part of reality (by providing information about the underlying methods and assumptions used in collecting and structuring them), and (3) under what conditions they can be used;
2. What extra resources and efforts are required to automatically combine transformed repositories with partially overlapping content domains into multi-modal data repositories;
3. What are the shortcomings in data and knowledge collection, storage management and analysis procedures and formalisms that prevent such transformation to be 100% successful?

Answers to these research questions intend to advance the state of the art in knowledge and data organization and to decrease the complexity for researchers by using ontologically coded content to act as a microscope and examine areas of the data very closely.

2.2 Background

2.2.1 Incentive

At a workshop sponsored by the International RDC/TMD Consortium Network and held at the IADR meeting in Toronto (July, 2008), current data regarding the reliability and validity of the RDC/TMD were presented based on an NIDCR-funded extensive multicenter study known as the 'RDC/TMD Validation Study'; presentation of these data was followed by invited critical commentary, and recommendations for revisions for the planned next RDC/TMD were suggested in order to improve evaluation and diagnosis of TMD. Based on the discussions at the workshop, a need for a consensus workshop to finalize the RDC/TMD version 2 emerged. In addition, the participants also indicated that there was a need to incorporate the RDC/TMD diagnostic taxonomy into a larger taxonomic framework that would include all of the other orofacial pain conditions in order to create a comprehensive orofacial pain taxonomy.

To initiate that effort, 'The International Consensus Workshop: Convergence on an Orofacial Pain Taxonomy', was held March 30 – April 1, 2009, Miami, Florida. The participants for the consensus meeting were selected so that these organizations and fields would be adequately represented:

- the International RDC/TMD Consortium Network of the International Association for Dental Research,

- the Orofacial Pain Special Interest Group of the International Association for the Study of Pain,
- the National Institute for Dental and Craniofacial Research,
- the American Academy of Orofacial Pain,
- the European Academy of Craniomandibular Disorders,
- the International Headache Society,
- patient advocacy, and
- the related areas of neurology, psychology, radiology, rheumatology and ontology, the latter represented by the three Directors of the Ontology Research Group, University at Buffalo, including the PI of this proposed effort.

The overall goals of the consensus workshop was to create Clinical Diagnostic Criteria for TMD (CDC/TMD), based on revisions of the RDC/TMD, for immediate clinical implementation and an initial draft of RDC for selected other orofacial pain conditions (RDC/OFP) where existing data are sufficient to identify draft criteria. It was decided that an adequate treatment of the ***ontology*** of pain in general, and orofacial pain in particular, together with an appropriate ***terminology***, is mandatory to advance the state of the art in diagnosis, treatment and prevention. The following consecutive steps were proposed [30]:

1. study the terminology and ontology of pain as currently defined,
2. find ways to make individual data collections more useful for international research,
3. develop an ontology for integrating knowledge and data over all the known basic and clinical science domains concerning TMD and its relationship to complex disorders,
4. expand this ontology to cover all pain-related disorders.

The effort conducted during the project reported on here involves mainly step 1 and 2.

2.2.2 Preliminary studies

2.2.2.1 Terminology and ontology of disease and disease perception

Many existing biomedical terminology standards rest on incomplete, inconsistent or confused accounts of basic terms pertaining to diseases, diagnoses, and clinical phenotypes. In [31], we outlined a terminological and ontological framework that encompasses diseases, their causes and manifestations, and diagnostic acts and other entities pertaining to the ways diseases are recognized and interpreted in the clinic. Inspection reveals that such entities have thus far not been adequately treated in standard vocabulary resources.

The National Cancer Institute Thesaurus (NCIT), for example, identifies '*Chronic Phase of Disease*' as a subtype of '*Finding'*, which it defines as: '*Objective evidence of disease perceptible to the examining physician (sign) and subjective evidence of disease perceived by the patient (symptom)*' [32].

This definition implies, however, that a disease does not exist except as one or other form of evidence. It thus illustrates a common conflation between processes on the side of the organism and the evidence for the existence of such processes. That this conflation is problematic is revealed when we need to link observable clinical phenomena to hypothesized unobservable biological causes.

A misplaced focus on observables is reflected also in the traditional practice of classifying diseases on the basis of patterns of similarities in signs and symptoms. This practice creates problems in face of the wide variations in clinical presentations of many diseases [33] and of the increasing importance for our understanding of the ways disease correlates with genetic and environmental variables [34]. The effective study of such correlations requires clinical research to be applied to ever larger pools of subjects drawn from geographically separated populations in multi-institution studies, requiring that the healthcare institutions involved embrace common standardized terminologies in capturing and sharing their data.

The approach we followed, designed to provide the resources in terminology and disease classification to support such standardization, rests on an account of diseases as dispositions rooted in physical disorders in the organism and realized in pathological processes. This approach helps us to do justice

1. to the existence of pre-clinical manifestations of disease (disorders can exist before they are realized in overt pathological processes);
2. to the combinations of disease and predispositions to disease which can exist within a single patient (as when an instance of disease of type A in a given patient is a risk factor for a second disease of type B); and
3. to the fact that the disease course and the clinical picture may vary widely between patients who have the same disease.

The central view adhered to any phenomenon that is standardly called '*clinically abnormal*' for a person (or any organism in general), is that it:

- is not part of the life plan for an organism of the relevant type (unlike aging or pregnancy),
- is causally linked to an elevated risk either of pain or other feelings of illness, or of death or dysfunction, and
- is such that the elevated risk exceeds a certain threshold level [35].

This treatment of 'abnormal' is distinct from those statistical treatments which do not take account of the overlap in the distribution of test results between normal and abnormal populations or of normal distribution extremes. What are standardly called 'normal variants' (for example a left lung with three lobes) do not satisfy criteria (2) and (3).

This work resulted in the publicly available 'Ontology for General Medical Science' [36].

2.2.2.2 Terminology and ontology of pain

In [37] we pointed out that the definition of pain issued by the IASP – '*an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage*' - ascribes a common phenomenology ('*unpleasant sensory and emotional experience*') to all instances of pain, together with the recognition of three distinct subtypes of pain involving, respectively:

1. actual tissue damage,
2. what is called 'potential tissue damage',
3. a description involving reference to tissue damage.

Clause 3 may be interpreted to mean that a mere description of a certain sort provides sufficient evidence that pain is present. The intent, as we understand it, is to assign those patient reports of pain that are not sufficiently grounded in observable manifestations of tissue damage to some other (for example psychological) realm. Problems arise, then, in the classification of cases of malingering. (Example: a patient presenting with pain and associated tissue damage was prescribed pain relief medication, and while moderate tissue damage remains the medication is effective and there is no longer pain. Because the patient has become addicted, he claims that there is still pain in order to obtain more medication.) Such cases are not pain, yet they will often be so classified by the clinician.

In line with [31], we thus pursue a view of pain as resting in every case on some physical basis perhaps as yet unknown. When, for example, there is a persistent pain in a patient's left temporomandibular joint (TMJ), then this is because some physical structure or substance in the organism is disordered (for example, the TMJ is deformed because of arthritis, or that part of the somatosensory cortex that serves as the projection of the left TMJ is disordered). As a result of this disorder, the organism acts in a certain abnormal way.

By 'physical basis' we understand any configuration of one or more physical components within the organism at any level of granularity, from a single nucleotide to an arthritically deformed joint. Where they are non-disordered – which means: such as to reflect the coordinated expression of the corresponding structural genes for an organism of the given type [18] – such configurations support those dispositions in the organism which are realized (manifested) in normal (= ordered) functioning. Where disordered, such configurations support dispositions to abnormal functioning, one family of which is manifested in experiences of pain. 'Symptom', as we here use this term, covers a restricted family of phenomena (including pain, nausea, anger, drowsiness), which are of their nature experienced in the first person.

Our goal here is to initiate the development of an approach which allows the clinician or researcher better to understand the physical basis underlying a report of pain and not just to stay at the level of reports and of the assumption according to which, if the patient says that it is pain (within the limits of language relating to tissue damage of one sort or another), then therefore it is pain (or as pain-clinicians will often say for the benefit of patients, 'all pain is real'). If the clinician expects concordance between stated intensity (the symptom) and the clinical findings (the signs), then significant problems will ensue, either in the form of dismissing the disorder, or in labeling the patient as 'psychiatric'. If, in contrast, the clinician understands the neuropathic and other non-peripherally localized contributions to pain experience, then this may serve a more adequate diagnosis.

We therefore define first what we shall call 'pain with concordant tissue damage', which we hold to be the canonical (normal, prototypical) and evolutionarily most basic case of pain, followed by a number of variant phenomena which are defined in terms of, and involve specific kinds of departures from, this canonical case. We then distinguish the following five different sorts of cases of pain and of pain-related phenomena (see Table 1):

PCT: pain with concordant tissue damage: the patient experiences pain of the evolutionarily most basic sort, which is to say: pain in response to and in concordance with tissue damage;

PNT: pain with peripheral trauma but discordant (elevated) relative to tissue damage: there is peripheral trauma, but the patient is experiencing pain of an intensity that is discordant therewith;

NN: neuropathic nociception: there is no peripheral trauma, but the patient is experiencing pain in result of a neuropathic disorder to the nociceptive system. An example is phantom limb

pain, where pain-system components in the brain which had been laid down through the PCT pain experiences activated earlier by tissue damage in the once present limb are re-activated.

In addition, we distinguish two related cases of non-pain-phenomena:

PBWP: pain behavior without pain: there is, for example, a mere report, and no pain is being experienced (a fact which may or may not be detectable by an external observer).

TWP: Tissue-damage without pain: tissue damage normally of the sort to cause pain does not activate the pain system.

In a full account, we would need to distinguish also various combination cases, for example where the patient experiences canonical (PCT) pain in conjunction with neuropathic nociception, as well as multiple subtypes, for example distinguishing acute and chronic pain varieties. In addition, we would need to take account of the fact that canonical The canonical has two basic temporal dimensions: subtype consists of pains of short duration: a cut, a local burn, an abrasion, is a brief duration stimulus and evokes a brief, intense experience of pain with accompanying reflex withdrawal that moves the body away from the stimulus. Following the injury there is a prolonged experience of usually less intense pain associated with inflammation that gradually recedes as healing occurs. The second is chronic pain, a long lasting sequence of experiences of pain, which may extend over many years without relief, and which may involve the patient visiting many specialists (ENT, headache, neurologist, TMD, psychologist) with no positive results. Our strategy is comparable to the way in which the results of genetic mutations or injuries affecting, for example, the human hand, are most effectively described in terms of specific kinds of departures from the anatomical structure of the normal human hand (with its five fingers, ten metacarpal bones, etc.). This strategy has been pioneered by the Foundational Model of Anatomy (FMA) Ontology, a scientifically well-established reference ontology of human (and more generally of mammalian) anatomy [18].

	Symp-tom	**Signs (= Objectively Observable Features)**	**Physical Basis**	**Examples**
CP: Canonical Pain				
PCT: Pain with Concordant Tissue Damage	Pain	Manifestation of tissue damage Report of pain concordant with stimulus sufficient to cause this tissue damage Protective response	Activation of nociceptive system through peripheral tissue damage	Primary sunburn Pain from strained muscle Pain from fracture Pulpitis
VP: Variant Pain				
PNT: pain with peripheral trauma but no concordant tissue damage	Pain	Report of pain associated with stimulus intensity insufficient to cause tissue damage	Activation of pain system through cognitive mechanisms regarding threat of tissue damage, the latter often based on peripheral non-nociceptive input to the CNS	Secondary sunburn without tissue damage Myofascial pain disorder Tension-type headache

				Chronic back pain
NN: neuropathic nociception (pain with no peripheral trauma)	Pain	Report of pain No identifiable pathological peripheral stimulus History of probable causes	Disordered nociceptive system Neuropathic (for example in result of demyelination of nerve fibers)	Trigeminal neuralgia Post-herpetic neuralgia Diabetic neuropathy
PRP: Pain-Related Phenomena Without Pain				
PBWP: pain behavior without pain		Sick role behaviors accompanied by normal clinical examination Report of pain discordant with physical signs Grossly exaggerated pain behaviors Identified external incentives	Description of pain relates to mental states such as anxiety, rather than peripheral tissue locus Misinterpretation of sensory signals by the emotional or cognitive systems Deception by patient	Factitious pain Malingering Anxiety-induced pain report
TWP: tissue-damage without pain		Manifestation of tissue damage normally of the sort to cause pain No reported pain	suppression of pain system by one or other mechanism	Stress associated with sudden emergencies Physiological damping of the pain process caused by adrenalin Placebo induced opioid analgesia Genetic insensitivity to pain

Table 1: Types of Pain and of Pain-Related Phenomena

2.3 Specific aims

The overall purpose of our project here is to develop an ontology which allows us to integrate datasets concerning patients suffering from various sorts of pain with the goal to obtain better insight in the complexity of pain disorders, specifically concerning the assessment of pain-related disablement and its association with mental health and quality of life. We will demonstrate the usefulness of this '*Ontology for pain-related disablement, mental health and quality of life*' (OPMQoL) by applying it to merge five existing datasets, collected independently from each other, containing data about patients with different sorts of orofacial pain.

Our goal is thus to build a realism-based ontology that makes it possible to describe the datasets in a uniform and formal way, and that is general enough to include other datasets in the same domain once they become available. The importance of this endeavor lays in its contribution to solving an important problem, namely that the phenotype of all orofacial pain conditions is insufficiently defined in terms of the scope, the natural history and/or clinical course of the disease subgroup of interest, and, most importantly, with respect to disease traits for which laboratory research has provided important pathogenetic insight [38]. The main clinical question that we will be able to answer by merging these datasets is how patho-anatomy and pathophysiology – in this case in the context of TMD - have an impact on pain-related disablement and quality of life.

We will build the ontology following the principles adhered to in the Open Biomedical Ontology Foundry (OBO-Foundry) [39], using Basic Formal Ontology (BFO) [40], and Referent Tracking (RT) [41] as generic semantic technologies. Whereas BFO provides facilities to describe what is general in reality, RT consists of mechanisms that allow data repositories to benefit maximally from ontologies. We will further resort to domain ontologies that have been developed in the same spirit such as the Foundational Model of Anatomy (FMA) [18] and the Ontology for General Medical Science (OGMS) [31]. By working in collaboration with *The International RDC/TMD Consortium Network* [42], and the *Orofacial Pain Special Interest Group* of the International Association for the Study of Pain [43], we intend this ontology to become a standard in the domain.

Our goal translates into the following specific aims:

- Aim 1: describe the portions of reality covered by the five datasets and acquire broad consensus in the field with respect to its face validity.
- Aim 2: design the bridging axioms required to express the data dictionaries of the datasets in terms of the OPMQoL ontology and translate these axioms in the query languages used by the underlying databases.
- Aim 3: validate the ontology by querying the datasets with and without using the ontology and by comparing the results in function of the clinical question identified.
- Aim 4: document the development and validation approach in a way that other groups can re-use and expand OPMQoL, and use our approach in other domains.

2.4 Significance and relevance to health

2.4.1 Assessment of quality of life and disablement.

The consequences of a disease include functional limitation and psychosocial disability. These two concepts refer to the individual's experience of limitations in function associated with the affected part of the body and to disarray in one's life, respectively. Models of disability emphasize the individual's self-report in describing these states and the centrality of these concepts as part of the disease and illness process. However, assessment approaches typically used in medicine and dentistry do not yet routinely include these domains. Yet, whatever the underlying disorder, they are both necessary and challenging.

Patients with musculoskeletal disorders in particular exhibit difficulties in functioning that range from temporary to persistent, from a mild state to a severe state, and from a particular isolated function of the involved joint to affecting the individual's quality of life. These problems in functioning are collectively referred to as *disablement*, where *disability* refers to only one aspect of disablement. The statistics regarding the prevalence of problems in functioning associated

with musculoskeletal disorders depend on (at least) the following five factors: spectrum of diagnoses, sampling frame with regards to community versus clinic, chronicity, selected level within the disablement model, and disability subtype [44]. The extent to which disablement affects individuals with musculoskeletal disorders is, as a result, not easily summarized.

The challenges are multiple. There is no shortage in available instruments for assessing functional limitation, psychosocial disability, and quality of life, each one of them being characterized, however, by strengths and weaknesses, either in general, or for specific sorts of disablement and/or in relation to specific disorders. The selection of an instrument depends furthermore on cultural matters, habits and/or legislation in different jurisdictions. There are issues of calibration such that questions whether treatments that improve structure improve impairment, and if not, whether it is because the clinical measures of function are inadequate, are not easy to answer. What would be *better*? How should *normal* be defined? What constructs (and instruments, if available) should a minimal assessment protocol be comprised of for use in clinical assessment and treatment of specific disorders?

Two classification models of disablement have been developed by the World Health Organization (WHO). *The International Classification of Impairments, Disability, and Handicaps (ICIDH)* uses the terms (1.) disease or disorder, (2.) impairment, (3.) disability, and (4.) handicap to describe four levels of disablement, where the overall emphasis was on the handicap, i.e. what the person could not do [45]. The second one is the *International Classification of Functioning, Disability, and Health (ICF)* that uses terms emphasizing health rather than disease, and which depicts disablement as dependent upon the interconnections of (1.) body function and structure, (2.) activity and participation, (3.) personal factors, and (4.) environmental factors [46].

A third widespread model is that of the Institute of Medicine (IOM) in which the respective disablement levels include (1.) pathophysiology, (2.) impairment, (3.) functional limitation, and (4.) psychosocial disability [47]. Both ICIDH and IOM sets of terms are in widespread use.

2.4.2 Pain, mental health and Quality of Life

Pain is defined by the International Association for the Study of Pain (IASP) as ‘an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage’ [48]. This definition has proved to be of considerable value, having led to 50 years of highly productive fundamental research on pain. On the other hand it has certain problems, as recently reflected by significant discussion by an IASP task force [49].

The study of chronic pain in humans needs to address complex issues of pain appraisal and response, which vary considerably from patient to patient, including the involvement of the emotional-affective system, cognitions, learning principles, pain behavior and societal and environmental factors [50]. Population-based studies to determine how people cope with chronic pain have shown that although a large majority of patients reporting chronic pain conditions maintain their lives without significant disruption or disability, large groups of patients are rendered dysfunctional by chronic pain [51]. Chronic pain dysfunction takes a personal toll in terms of emotional suffering and a societal toll in terms of disproportionately elevating the costs of health care. Efforts to help restore dysfunctional chronic pain patients to functional lifestyles are extremely difficult [52]. All symptoms, including physical symptoms such as pain, arise from the interaction of multiple factors - genetic, developmental, environmental - as they encounter precipitating events. When the events are negative, or interpreted as negative, the resolution of this dynamic process is distress: a dynamic, dysphoric organismic state which is noxious or aversive and from which relief is sought. Specifically, relief is sought from the experience of

negative or aversive symptoms in the form of attempts to cope which may be adaptive or maladaptive.

2.4.3 Temporomandibular disorders and Quality of Life

Temporomandibular disorders (TMD) represent a set of conditions that affect the masticatory muscles and the temporomandibular joint (TMJ), either isolated or in combination. TMDs are characterized by pain and mechanical limitations, can range from simple to complex and exhibit substantial symptom overlap with disorders affecting other parts of the body.

The Technology Advancement Conference by the National Institutes of Health on TMD [53] defined these disorders according to two broad aspects: pain and psychosocial dysfunction. There now seems to be increasing evidence that these two aspects are the important, if not cardinal, features that make patients seek treatment. In most cases the diagnosis of TMD is based on careful patient history taking and clinical examination, which depends on patient report of levels of pain and discomfort of the TMJ and associated muscles. Often patients with TMD also describe symptoms of pain and dysfunction affecting ears, eyes and/or throat and headaches that involve some or all of the frontal, temporal, parietal, occipital and neck regions. Clinical examination methods include measures of quasi-objective factors that define limitations of mandibular function and tenderness of head and neck muscles. These are currently based on a consensus among leading researchers and clinicians internationally [54].

Probably the most widely studied measure of these variables is the Research Diagnostic Criteria for TMD (RDC/TMD) developed at the University of Washington [55]. This system has two assessment components. Axis I, a clinical and radiographic assessment, is designed to differentiate myofascial pain, disc displacement, and arthralgia, arthritis, and arthrosis. Axis II evaluates psychological status and pain-related disability. Numerous publications have suggested aspects of the RDC/TMD that could be improved to more effectively distinguish TMD cases from controls and differentiate diagnostic subgroups [56]. The first aim of the RDC/TMD Validation Project, a project that led to two datasets used in our proposed effort, was to rigorously establish the reliability and validity of the RDC/TMD diagnostic protocol in its published form. The second aim was to propose modifications for the protocol that would improve its reliability and validity as a taxonomic system.

In the case of temporomandibular disorders, the amount of patients that are rendered dysfunctional by chronic pain comprises 20-30% of the clinic population [51]. Examining the inter-relationships between somatization, emotional status and pain dysfunction shows that functional and dysfunctional TMD patients differ significantly in the extent of their depression and their levels of somatization: dysfunctional chronic pain patients are significantly more depressed, and they report having significantly more nonspecific physical symptoms than functional TMD patients [51]. Similarly, somatization influences clinical findings where the clinical examination incorporates subjective responses. TMD patients scoring high in self-reported presence of nonspecific physical symptoms such as palpations, trembling and dizziness also report significantly more muscles tender to palpation on clinical examination. The current perspective regarding TMD is now multidimensional, with an appreciation that a combination of physical, psychological and social factors may contribute to the overall presentation of this disorder – hence the preference for a biopsychosocial integrated approach [54]. Because several studies have reported that musculoskeletal disorders of the stomatognathic system resemble musculoskeletal disorders and pain disorders in general, e.g. [57], it can be expected that the insights that will be obtained by the effort proposed here will not only benefit TMD patients, but patients with other pain-related disorders as well.

3 Study materials

We obtained six datasets (one more than proposed) – we named them for further reference according to their origin – which cover roughly the same domain, but are distinct in various respects. Although a number of variables are identical, variables involving, for instance, somatization, depression and anxiety, are different because measured with distinct assessment instruments. Finally, there are also some cultural influences related to pain report that have an impact on the comparability of the five data sets, despite the use of common instruments.

3.1 The '*German Dataset*'

This study set – received with supporting documents May 14, 2011 – was collected from 390 patients seeking treatment for orofacial pain at the Department of Prosthodontics and the Department of Prosthodontics and Materials Sciences of the Universities of Halle and Leipzig, respectively [58]. The inclusion criterion was that patients should have had at least one diagnosis in accordance with the German version of the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) [55, 59]. The goal of the study was to characterize the level of impairment of oral health-related quality of life. Patients could attend at their own initiative or were referred by their dentist, physician, or physiotherapist. Inclusion criteria were that patients had at least one diagnosis according to the German version of the RDC/TMD which is almost identical to the English-language original. The only difference is that depression and somatization are assessed according to recommendations of the working group on pain assessment of the German Chapter of the International Association for the Study of Pain.

The 'Allgemeine Depressionsskala' with 20-items, the German translation of the Center for Epidemiological Studies Depression Scale (CES-D), was used to assess depression whilst the 'Beschwerdenliste', a 24-item instrument, was used to assess somatization. For these instruments population-based normative data are available which allow the classification of 'normal', 'moderate', and 'severe' depression or somatization – the categorization recommended by the original English-language RDC/TMD. For some of the subjects, depression was assessed using the 'Gießen-Test' with 6 items, another well-accepted instrument in Germany which is able to assess depression based on population-based norms allowing the RDC/TMD-recommended categorization. Some patients had missing data for depression (n=57, 13.7%) and somatization (n=5, 1.2%). Only subjects with TMD pain in the prior 6 month to this study filled in the Graded Chronic Pain Scale (N=301).

Oral health-related quality of life (OHRQoL) was measured using OHIP-G, the German version of the Oral Health Impact Profile. The purpose of the (OHIP) is to provide a comprehensive measure of self-reported dysfunction, discomfort and disability arising from oral conditions with the goal to assess the social impact of oral disorders [60]. It is based on an adaptation by Slade and Spencer of the World Health Organization's Classification of Impairments, Disabilities and Handicaps. In this model, impacts are organized linearly to move from a biological over a behavioral to a social level of analysis. Slade and Spencer adapted this by proposing seven dimensions of impact of oral condition, each dimension being assessed from 49 questions on the type of problems experienced. The OHIP-G has 49 items derived from the English-language OHIP and four items specific for the German population. For each OHIP question, subjects were asked how frequently they had experienced the impact in the last month.

The study set came with a codebook consisting of 161 variables and a technical report explaining certain dependencies and implicit assumptions related to the RDC/TMD section of the dataset [61].

3.2 The '*UK2 dataset*'

The existence of this dataset was unknown at the start of the project and processing of it not foreseen. It became available Sept 12, 2012, thanks to contacts with our European collaborators. The dataset contains data about the 2298 participants from a UK population study of orofacial pain and came with the data collection questionnaires for the population study and examination attached as structured history.

3.3 The '*Swedish Dataset*'

This set – received on Nov 9, 2012 – contains data about 46 consecutive Atypical Odontalgia (AO) patients recruited from 4 orofacial pain clinics in Sweden as well as data about age- and gender-matched control patients, 35 of which being painless and 41 being TMD patients. The AO group had pain located in a region where a tooth had been endodontically or surgically treated, chronic pain of at least 6 months duration, and pain with no pathological cause detectable in clinical and radiological examinations. Painless controls were routine dental patients which had tooth extractions (trigeminal nerve damage). TMD patients had pain during the last month and a TMD Axis I diagnosis with pain from the RDC/TMD. Clinical measures assessed include pain location, number of teeth and root fillings, mandibular range of motion variables, and Pressure Pain Threshold (PPT). All AO patients and controls underwent a neurological and somatosensory examination. Standardized quantitative sensory testing (QST) included the mechanical detection threshold (MDT), the mechanical pain threshold (pinprick) (MPT), dynamic mechanical allodynia [DMA (brush)] and dynamic mechanical allodynia [DMA (vibration)], wind-up ratio (WUR), thermal thresholds (for warmth, cold and heat pain). Self-report measures involved general patient characteristics and pain characteristics, a Swedish version of the short-form McGill pain, the Jaw function limitation scale (JFLS), and a shorter version of the symptom checklist-90 (SCL-90) according to the RDC/TMD. Quality of life was measured using the SF 36.

3.4 The '*Hadassah Dataset*'

This dataset (306 patients) was collected at the Orofacial Pain Clinic at the Faculty of Dentistry, Hadassah, The Hebrew University, Israel [62]. All consecutive patients (n = 328) visiting the clinic between 2005 and 2007 were interviewed at the first visit before medications were prescribed. The resultant data, including a standard pain history, was recorded on an intake form. Patients were asked to rate pain duration, quality, and average pain intensity over the previous week. Pain quality was assessed by asking the patients to choose one or more of the following descriptive terms: electrical, stabbing, throbbing, pressure, burning, or any combination of the five terms. Pain intensity was rated by a verbal pain scale (VPS). Pain that began following a clear traumatic event was defined as "posttraumatic" and as "primary" in the absence of a traumatic onset.

Inclusion criteria comprised a complaint of persistent facial pain, which may also involve the head. "Persistent" refers to pain that was present for a minimum period of 3 months. Headache and various sorts of facial pain were diagnosed by means of the criteria published by the International Headache Society. Painful temporomandibular disorders were diagnosed according to the criteria published by the American Academy for Orofacial Pain and the RDC/TMD. Patients with rare diagnoses (n = 22) were excluded so that the final group consisted of 306 patients. Depression, anxiety, and cognitive rumination were in contrast to the other two datasets, not assessed in this study.

The set came as an Excel spreadsheet without data dictionary which had, as a consequence, be developed in collaboration with the source during this project and was completed Oct 1, 2013.

3.5 The '*UK1 dataset*'

This dataset involves 92 out of the in total 168 British patients with facial pain of non-dental origin is collected on the basis of a throughput of some 600 new patients and 800 review patients per year. Criteria for inclusion are facial pain of non dental origin present for a minimum of three months. The data is collected on structured history sheets and the following psychometrically tested questionnaires used in many trials are completed by patients prior to their first visit : Brief Pain Inventory, Hospital Anxiety and Depression scale, Graded Pain Chronic Scale of Von Korff, McGill Pain Questionnaire, and Pain Catastrophising questionnaire. Diagnosis is made using the IHS criteria, not the full RDC. Around 60% of the patients have a TMD diagnosis and we have a large proportion of trigeminal neuropathic pain, especially trigeminal neuralgia (the largest collection in the UK).

It was found not to be correctly structured and we worked with the UK1 team to develop an appropriate database who then inputted the data in the correct format in order to make it useable and potentially comparable with the other datasets. The dataset became available Jan 2, 2014.

3.6 The '*US dataset*'

This dataset (724 patients) resulted from a collaboration effort of the University of Minnesota, the University at Buffalo and the University of Washington and was collected during the NIH funded *RDC/TMD Validation Project* (U01-DE013331) carried out to validate, and, when needed change, '*The Research Diagnostic Criteria for Temporomandibular Disorders*' (RDC/TMD) [55]. Beginning in August 2003, study participants were consecutively recruited until three-fourths of the study sample had been enrolled [63]. At this point, it was necessary to institute selective recruitment in order to fill out the recruitment goals for the less common TMD diagnoses. Other subgroups requiring selective recruitment were older age categories for normal participants and TMD pain cases needed for completing Axis II studies. Selective recruitment was continued until study closure in September 2006. Participants were drawn from 2 sources: direct referrals from local health care providers to the respective university-based TMD centers (i.e., clinic cases) and responses to community advertisements (i.e., community controls and cases). Recruitment was designed to include cases with a full spectrum of TMD signs and symptoms. Participants, ages 18 to 70 years old, entered the study as putative TMD cases or controls based on the inclusion and exclusion criteria. The inclusion criteria for study eligibility differed from the published RDC/TMD diagnostic criteria by assigning putative case status to individuals who reported a minimum of 1 of the 3 cardinal symptoms of TMD: jaw pain, limited mouth opening, or temporomandibular joint (TMJ) noise. Participants who denied currently having any of these symptoms were enrolled as controls. IRB approval was obtained at each of the 3 study sites prior to initiating the Validation Project.

First parts of the dataset for axis 2 were received Oct 12, 2012, the part with sufficiently elaborated data dictionary for axis 1 on Jan 2014 but with distinct case-pseudonyms. Aligned case files were received June 17, 2014.

4 Foundations of the research conducted

In the domain of healthcare information technology (HIT) it has been commonly accepted for some years now that both the development and use of clinical terminology should be supported by formal methods. Although this is a thesis that we strongly support, we wish no less strongly to insist that formal methods alone are not enough. The use of a Description Logic-based system appears, for example, not to have provided any guarantee for the absence of errors in SNOMED-CT [64], one of the most popular formal biomedical terminologies today.

4.1 Ontology

The word 'ontology' - as mass noun - was originally used to denote a philosophical discipline devoted to the study of what entities exist in reality and how these entities relate to each other. Within that context, the word is sometimes also used as a count noun to refer to one's account for how reality is structured, thus allowing statements such as 'Aristotle's ontology differs from Plato's ontology'. It is also as count noun that the word 'ontologies' became popular in computer science in general and biomedical informatics in particular, but then in the meaning of representational artifacts of various sorts each one describing some part of a domain relevant for a particular purpose. Examples of what nowadays are claimed to be ontologies are controlled vocabularies, nomenclatures, terminologies (formal or not), and also classification systems. It is however unfortunate that most authors of these artifacts lack a background in the discipline of ontology itself, and this often on top of insufficient insight in terminological principles [65] and in the semantics of the representation language they use [66]. It is as a result very often unclear what - if anything at all - the representational units (usually terms) of these artifacts actually represent, and to what degree the structural organization of these units corresponds to how reality is organized in contrast to our perception thereof, or the way we talk about reality [2].

4.2 Realist Ontology

With the extremely positive response to the creation of the Open Biomedical Ontologies (OBO) Foundry [39] it became clear that a role had to be played by *realist ontology* in making better biomedical terminologies. Realist ontology helped in detecting errors and in ensuring intuitive principles for the creation and maintenance of systems of a sort that can help to prevent errors in the future. More importantly still, however, it helps in ensuring that terminologies are compatible with each other. Note that we say '*realist ontology*', in order to distinguish ontology in our understanding from the various related artifacts [67] which go by this term in contexts such as formal knowledge representation. It is a realist conception of ontology which underlies statements such as:

The UMLS is an extensive source of biomedical concepts. It also provides a large number of inter-concept relationships and qualifies for a source of semantic spaces in the biomedical domain. However, the organization of knowledge in the UMLS is not principled nor consistent enough for it to qualify as an ontology of the biomedical domain [68].

In the tradition of analytical philosophy, ontology is understood by the OBO Foundry community not as a software implementation or as a controlled vocabulary, but rather as '*the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality*' [40]. Ontology as it concerns us here is a theory of those higher-level categories which structure the biomedical domain, the representation of which needs to be both unified and fully coherent – and as closely allied as possible to the representations used by clinicians in formulating patient data – if terminologies and coding systems are to have the

requisite degree and type of interoperability. Ontology in this realist sense has successfully been used as a method to find inconsistencies in terminologies and clinical knowledge representations. The method has also proved useful in drawing attention to certain problematic features of the HL7 RIM [69].

One of the major insights brought about by realist ontology in the healthcare domain is that biomedical terminologies can only be compared amongst each other, or used without loss of information within data repositories such as electronic healthcare records or clinical trial databases, if they share a common framework of top-level ontological categories [70]. Often one talks in this connection merely of a shared or common *semantics*, meaning thereby the sort of regimentation that can be ensured through the use of enabling technologies such as OWL [71] that currently enjoys a wide interest through its association with the Semantic Web project, not to forget systems such as Protégé that are able to cope with them in a user-friendly way.

On closer inspection, however, one discovers that the 'semantics' which comes with languages like OWL is restricted to that sort of specification of meaning that can be effected using the formal technique of mathematical *model theory*, which is to say that meanings are specified by associating with the terms and sentences of a language certain abstract set-theoretic structures, taking Alfred Tarski's 'semantic' definition of truth for artificial languages as paradigm [72]. But model theory is metaphysically and ontologically almost completely neutral. Merely to formulate statements in a language such as OWL is far from building an ontology in the sense of ontology that is employed by analytical philosophers, and neither would translating a terminology into OWL turn it into an ontology. Such translation would indeed allow consistent reasoning about the 'world' – but only in the model-theoretic sense of 'world' that signifies not the flesh-and-blood reality with which biomedicine is concerned, but rather merely only some highly simplified set-theoretic surrogate. The task of ensuring that the latter somehow corresponds in broad terms to the real world of what happens and is the case, was in the semantics biomedical literature almost never addressed. Now it has become clear that the whole detour via semantic models is in fact superfluous: the job of ontology is not the construction of simplified models; rather, a biomedical ontology should directly correspond to reality itself in a manner that maximizes descriptive adequacy within the constraints of formal rigor and computational usefulness. Applying realist ontology to terminologies and data collections, even the assessment instruments through which data are collected, means in the first place applying it to those entities in reality to which these artifacts of the human intellect refer, such as concrete patients, diseases and therapies. We do this to serve at least one important goal, namely making terminologies coherent, both internally as well as in their relation to the databases in or for which they are used.

European and international efforts towards standardization of biomedical terminology and electronic healthcare records were focused over the last 15 years primarily on syntax. Semantic standardization was restricted to terminological issues around the semantic triangle paradigm [13] on the one hand and to issues pertaining to knowledge representation (and resting primarily on the application of set-theoretic model theory) on the other hand [73]. Moves in these directions are indeed required, and the results obtained thus far are of value both for the advance of science and for the concrete use of healthcare telematic applications. We can safely say that the syntactical issues are now resolved and also that the problems relating to biomedical terminology (polysemy, synonymy, cross-mapping of terminologies, …) are well understood – at least in the community of specialized researchers. Now, however, it is time to solve these problems by using the theories and tools that have been developed so far, and that have been tested under laboratory conditions. This means using the right sort of ontology, i.e. an ontology that is able explicitly and unambiguously to relate coding systems, biomedical

terminologies and electronic health care records (including their architecture) to the real world [74].

The message of realist ontology is that, while there are various different views of the world, this world itself is one and unique. It is our belief that it is only through that world that the various different views can be compared and made compatible. To allow clinical data registered in databases by means of coding (and/or classification) systems to be used for further automated processing, it should be crystal clear whether entities in the coding system refer to diseases or rather to statements made about diseases, or to procedures and observations, rather than statements about procedures or observations. As such, coding systems should be given a precise and formal semantics that is coherent with the semantics of the record as well as with the real world parts that are described by them.

4.3 Referent Tracking

Chronic disorders often lack strong association to clinical findings and patients suffering from them are typically cared by a multitude of providers from various disciplines because of co-morbid disorders. Consequently, phenotypic characterization is critical and very complex. This requires the use of data storage and analysis methods across multiple medical disciplines with the goals of (1.) adding additional analysis tools to the task of linking genotype and phenotype for immediate clinical research needs, and (2.) developing heuristics that will position investigators for further research using multiple existing datasets associated with identified co-morbid disorders as well as developing prospective studies.

Referent Tracking (RT) [75-78] is a methodology for data acquisition, storage and analysis based on Basic Formal Ontology (BFO). Ontology, as a scientific discipline, studies (1.) what entities exist in reality and (2.) how these entities relate to each other. BFO and RT, by combining ontology with computer science, help to distinguish various sorts of entities formally in ways that not only allow investigators to better use software programs, but also to let software programs discover new information autonomously.

Amongst ‘first-order entities’, BFO deals with what is generic (symptoms, disorders, treatments, guidelines and so forth), while RT deals with what is specific (e.g. that patient John Doe’s TMD is not the same as Joe Smith’s TMD, though both are instances of the generic disorder known as ‘TMD’). In contrast to prevailing paradigms, BFO and RT also deal with two kinds of ‘second-order entities’: (1) beliefs about first-order entities (hypotheses, diagnoses, …), and (2) representations (i.e., data) to document and communicate what is relevant. In addition, representations can be either about first-order entities directly or about second-order entities. Thus one can express using BFO that some forms of TMD are inflammatory disorders (relating first-order entities to each other) and that it might be caused by specific vulnerabilities, environmental exposures, or any combination (thus expressing a specific scientific theory). Similarly, RT allows to group dynamically at multiple levels of granularity patients with certain characteristics formally, or to compare different opinions about concrete cases organized not only on the basis of first-order characteristics but also second-order ones.

Both RT and BFO employ a formal theory to keep track of these distinctions between first- and second-order entities and between what is specific or generic throughout the history of that part of reality which the data are intended to represent (e.g., the time-course of the characteristics belonging to a specific individual participant). This allows, for instance, for a dynamic reclassification of patients in terms of the history of their disease at different time points or over different time periods, or in terms of new versions of terminology or classification systems that are introduced before, during or after data has been collected [15].

RT-compatible representations are not dependent on the context of a specific study, nor are they biased by the purpose(s) for which data collections are designed. Because of the ontological principles applied, data are rather organized in a way that mimics the structure of reality and optimized to detect in individuals the presence of patterns that deviate from what the scientific hypothesis suggests, even when both science and individuals are in flux. When used in combination, RT and BFO offer thus an ideal platform to integrate data from various studies in order to build data collections that are not only suitable to confirm or reject extant hypotheses, but also to assist in new hypothesis generation emerging directly from the structure of the data.

The general methodology for endeavors of this kind follows three steps.

The first step is an ontological analysis of the variables used in data repositories and assessment instruments used for data collection which results in a representation of the entities in reality about which the data are collected in terms of these variables. The distinction in such an approach is that the model (representation) conforms to formal rules and thus tests the fit of the data to the model rather than the other way around.

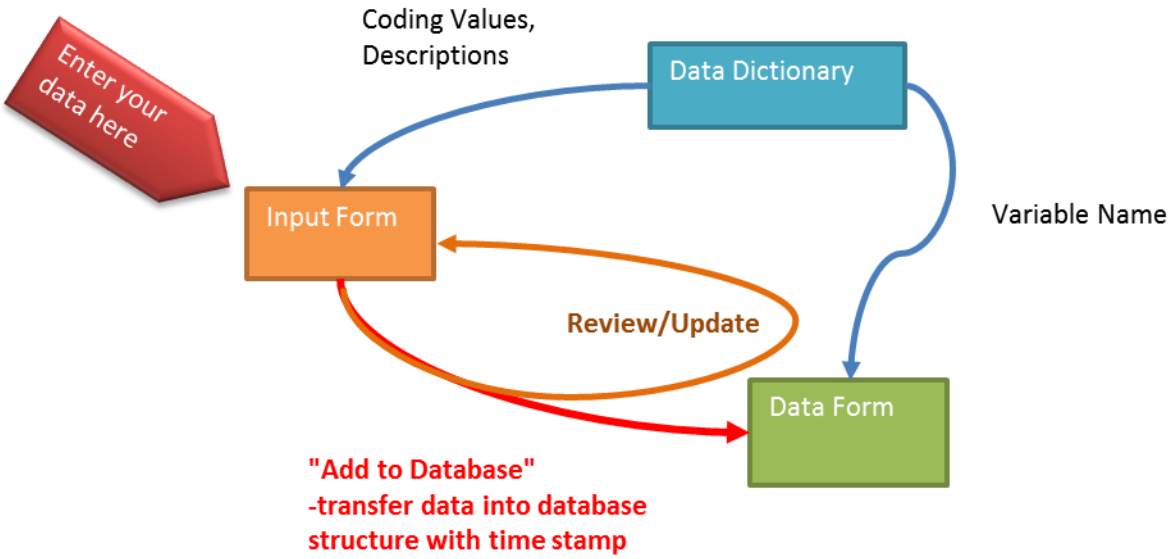
The second step is to study how the data elements that are currently used in ongoing studies or proposed in new ones line up with the data elements required to have a representation which is faithful to the reality as embedded into the data as observed – that is, the ontological principles are built into the repository. Such an analysis goes far beyond the mainstream approach towards common data elements that ignores faithfulness to reality. At this point we can make suggestions for improvements.

The third step is to build the overall structure of the repository followed by data population from existing repositories.

5 Development of a Data Collection/Entry Platform for UK1 dataset

5.1 Data Entry Form Worksheet Structures

a. Each questionnaire has 3 worksheets
   i. Input = Single column for data entry + utilities
   ii. Data = Database structure + time tracking
   iii. Data Dictionary = basis of input and data from



5.2 Input Form Features

1) Extract "Coding Values" from Data Dictionary
2) Generate "Drop-down box" for easy data entry
3) Pop-up flag about restriction of the cell by mouse click or arrow key
4) Allow "Add to database" features
5) Allow reviews/update on previous entered patient data

6 Data Collections and supporting documentation from an Information Artifact Ontology perspective[1]

The purpose of the work reported on here was 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].

6.1 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 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.

6.2 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).

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].

[1] This section is an update from work published as 4. Ceusters W. An Information Artifact Ontology Perspective on Data Collections and Associated Representational Artifacts. Medical Informatics Europe Conference (MIE 2012), Pisa, Italy, August 26-29, 2012, Stud Health Technol Inform. 2012;180:68-72.

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 a relation of *aboutness* to some *portion of reality* [4]
Representational Artifact (RA)	an **ICE** that is believed to *represent* a *portion of reality* external to the representation (modified from [5])
Representational Unit (RU)	an **RA** which, according to the structural conventions on the basis of which it is designed, is not built out of any other **RAs**
Denotator	an **RU** which *denotes* an *entity* (i.e. without providing a description) [6]
Term	an **RU** which is a general expression in some natural language used to refer to *portions of reality* (modified from [5])
Composite Representation	an **RA** built out of RAs as its parts (modified from [5])
Data Collection	a **composite representation** built out of **data items**
Data Dictionary	a **composite representation** describing, inter alia, what **data items** in a **data collection** are *about*, including a **data format specification**
Terminology	an **RA** consisting of **terms** (modified from [5])
Ontology	an **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 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	an **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 an **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

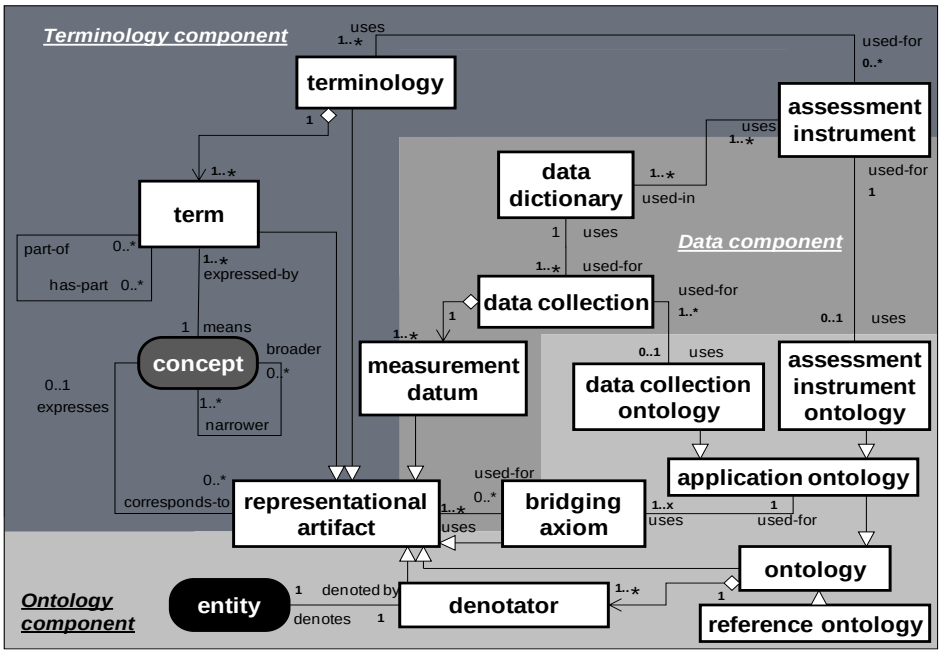


Figure 1: relationships amongst sources and their components.

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.

6.3 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

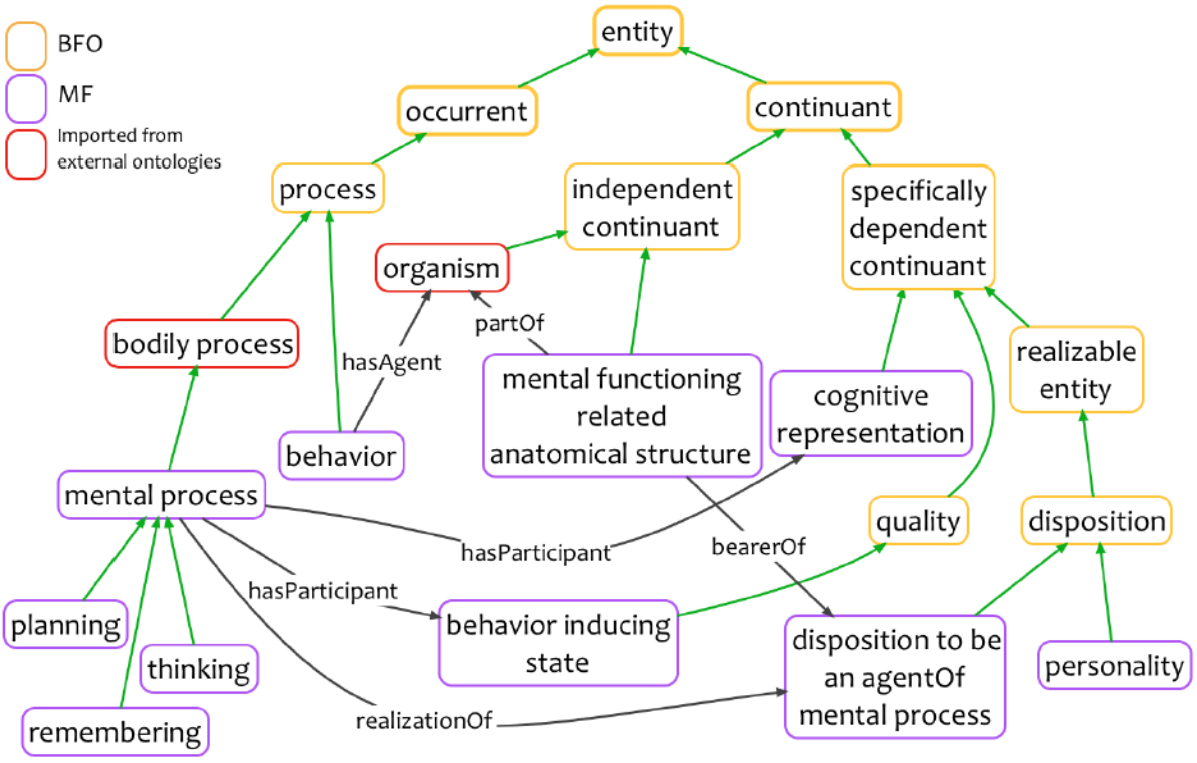
solution for applications that have to integrate/interface instances of both of these types of resources while still embracing Ontological Realism. Because, as proposed here, both *terms* (used in terminologies, assessment instruments and data dictionaries) and *denotators* (denoting particulars when 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].

7 Towards ontologies for mental functioning and unexplained syndromes

7.1 The mental functioning ontology

Affective science is the study of emotions and of affective phenomena such as moods, affects and bodily feelings. It combines the perspectives of many disciplines, such as neuroscience, psychology and philosophy. Emotions have a deep and profound influence on all aspects of human functioning, and altered or dysfunctional emotional responses are implicated in both the etiology and the symptomology of many pathological conditions.

Based on the Basic Formal Ontology [79] and on the foundations for an ontology of mental disease [80], and being developed in the context of the OBO Foundry [39], the Mental Functioning Ontology [81] is a modular domain ontology aiming to represent all aspects of mental functioning, including mental processes such as cognitive processes and qualities such as intelligence. MF grounds mental functioning entities in an upper level ontology, and gives a framework within which mental functioning can be related to ontological descriptions of related entities in other domains such as neuroanatomy and biochemistry. Modules of MF that are actively under development are those for cognition, perception and emotion.



7.2 Applying ontological principles to the analysis of unexplained syndromes

7.2.1 Materials and Methods

We performed as first step a literature review of papers about MUS in general and specific types thereof to identify the sorts of agreements and controversies requiring representation in our

ontology. We studied in a second step the principles of Ontological Realism [82], criticisms thereof [1], as well as suggested solutions [83]. and assessed in particular - in line with principle P3 (Table 1) - whether the Ontology of General Medical Science (OGMS, http://code.google.com/p/ogms/) which represents entities such as *disease*, *sign*, *symptom*, *clinical picture*, *diagnosis*, and so forth [31], and the Information Artifact Ontology (IAO) which provides an overarching perspective on entities such as *terminologies*, *classification systems*, and *diagnostic criteria* [84], could serve as feeder ontologies. We finally used the insight obtained in the second step to outline the conditions and ontology design criteria under which the issues identified in the first step can be resolved.

Table 1 - Main principles of Ontological Realism [82]

Principles for Reference Ontologies	
P1	***Reference ontology principle***: a reference ontology should cover the terminological content of the settled portions of a given scientific discipline, including only general terms which are assumed to denote corresponding universals in reality and assertions of certain relations between instances thereof.
P2	***Principle of consistency with established science****:* the assertions of which a reference ontology consists at any given stage should be consistent with the best available settled science that is current at that stage.
P3	***Principle of instantiation:*** a term should be included in a reference ontology only if there is experimental evidence that instances to which that term refers exist in reality.
P4	***Principle of asserted single inheritance****:* each reference ontology module should be built as an asserted mono-hierarchy.
Principles applying to any realism-based ontology	
P5	***Application ontology principle:*** in areas where research is still exploratory and results provisional, application ontologies are to be built as far as possible as extensions of corresponding reference ontologies.
P6	***Ontology path dependence principle***: decisions made by the creators of an ontology should as far as possible be made on the basis of the degree to which they advance the consistency of that ontology with the reference ontologies already existing in relevant domains.
P7	***Principle of Aristotelian definitions:*** any term 'A' asserted to have parent term 'B', should be defined as 'A=$_{def}$. a B which C', where 'C' expresses some condition on those instances of B which fall within the A's.
P8	***Principle of obsoletion:*** if a term in an ontology fails in designation, then it must immediately be obsoleted.

7.2.2 Results

Table 2 lists the areas in the domain of MUS for which scientists have thus far not yet reached agreement about what is going on in a patient which exhibits symptoms that are suggestive for MUS, or about how one should proceed to make a reliable diagnosis. These areas are thus not yet part of settled science and create challenges for a MUS reference ontology.

Table 2 - Challenges for a realism-based MUS ontology

Debates about the pathophysiological basis:	
C1	whether MUS form a subclass of somatoform disorders, are separate clinical syndromes, or no syndromes at all [85],
C2	whether patients with MUS have a pathology either (1) inside or (2) outside the brain alone, or (3) in both brain and other bodily structures simultaneously, or (4) have no pathology at all [86].
Problems with coherence of diagnostic criteria [87]:	
C3	frequently updated whereby some patients classified by means of an earlier version become classified differently later without there being any significant change in their disease course,
C4	criteria issued by distinct authors are such that the same patient would be classified differently depending on the criteria used,
C5	some criteria classify patients with very distinct phenotypes in the same category.
Diagnosis strongly based on symptom severity [88]:	
C6	patients and physicians have been found to be reluctant to entertain the idea of psychosocial factors resulting in not mentioning, exaggerating or down-playing symptom severity.

Table 3 and Table 4 list some design recommendations resp. representational units for a MUS ontology such that it satisfies the principles of Table 1 given the challenges identified in Table 2.

Table 3 - Design recommendations and conclusions for a realism-based MUS ontology

R1	'MUS', whether in the meaning of medically unexplained symptom or syndrome cannot be a representational unit in a reference ontology.
R2	The current OGMS definitions for *syndrome* and *sign* have shortcomings which would make it risky to define 'MUS' terms on their basis.
R3	The OGMS' representational units *clinical picture* and *diagnosis* are inspirational for defining similar classes relevant to MUS, but fall short in being encompassing.
R4	Assessment instruments and diagnostic criteria sets for MUS are to be analyzed as composite representations whose components are about universals.
R5	A MUS ontology should be an application ontology.

Table 4 - Foundational units for a MUS ontology

Clinical Representation =def. – A representational artifact of a phenotype that is inferred from the combination of laboratory, image and clinical findings about a given patient.
Unexplained Clinical Representation =def. – A clinical representation that when used as input for an interpretive process does not lead to a diagnosis.
Diagnosis of MUS =def. – A representation of the conclusion of an interpretive process that has as input an unexplained clinical representation of a given patient and as output an assertion to the effect that no diagnosis has been established.

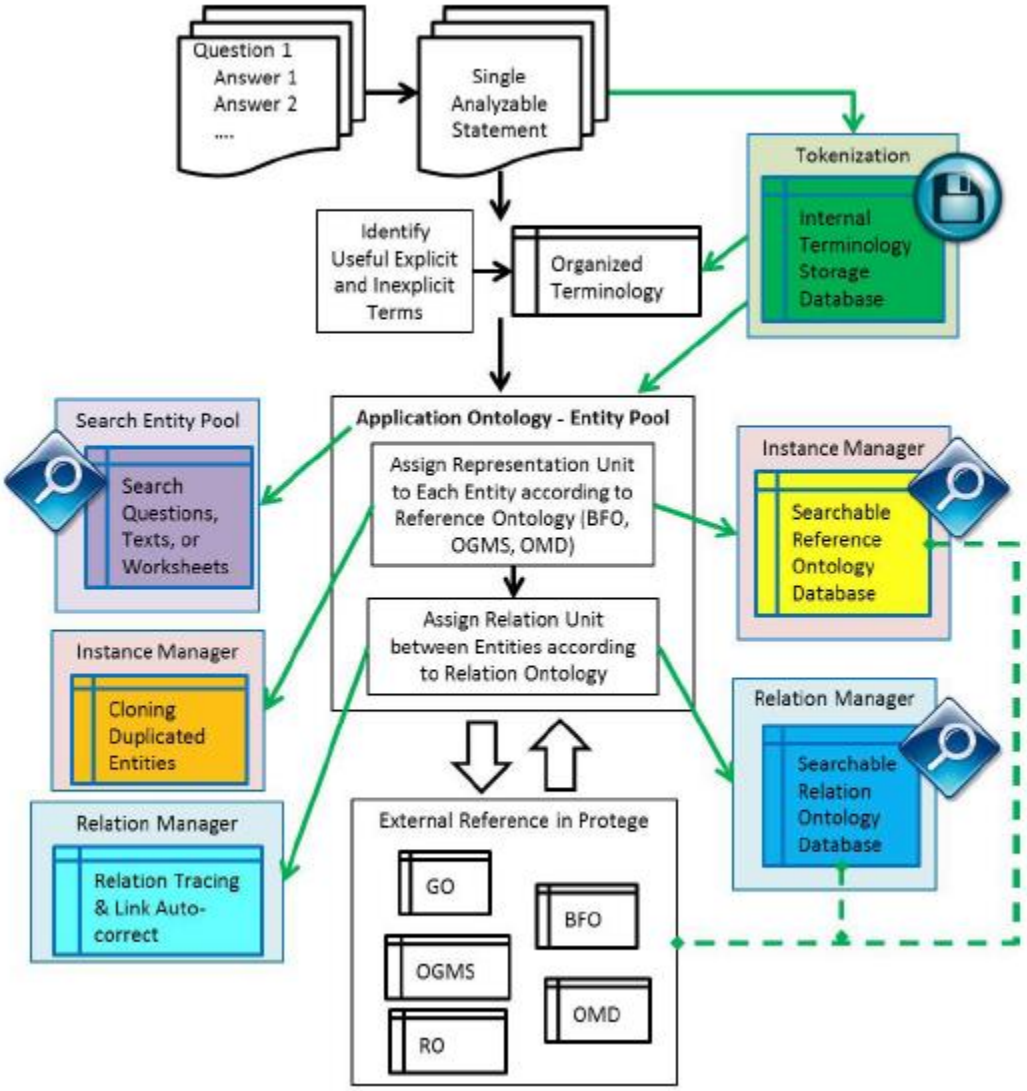
8 Conclusion

Our research did not reveal any indications that the principles of Ontological Realism make the latter inadequate for application to MUS. OGMS was however found to leave certain questions unanswered, most importantly the precise relationships between clinical phenotype and disorder [89]. This makes it difficult, if currently not impossible, for MUS experts to formulate hypotheses about the nature of MUS in terms of OGMS.

9 Ontological analysis of assessment instruments

9.1 Methodology

The goal of this part of the effort was to identify the number and types of particulars on the side of the patient that must exist for a specific answer to a questionnaire or assessment instrument question to be faithful to reality. To do so, a computer-assisted method for question/response analysis was developed and implemented in Excel's VBA using the Access database management system. This allowed, after copying the questions and possible answers to an Excel spreadsheet, to semi-automatically generate analysis templates for the questions and link the identified entities to an expanding ontology.



The central components that have been developed are:

OntologyTools: a Microsoft Excel Add-in (.xlam) programmed in Visual Basic Application (VBA). It is embedded as a form of Microsoft Excel Customized Ribbon Menu and programmed in XML. It controls multiple ontological analysis workbooks, manages connections between each analysis workbook and Ontology databases, and facilitates the processes of terminology alignment and application ontology development

Tokenizer/Instance Manager: a tool which calls out a UserForm/VBA Module with 2 major functions: Tokenizer and Instance Manager. Tokenizer can 1) execute the tokenization process of an analysable statement, 2) collect useful/exclusion terms into databases in OntologyDB, and 3) insert tokenized terms or tokens into a specified section under the statement. The tokenizer allows differentiation between 1) linguistic function and stop words (words such as "a", "the", "of" etc.), 2) punctuations and symbols, meaningful phrases, 4) inexplicit tokens amongst which 5) convertible inexplicit tokens. The tokens falling into category 1 and 2 are removed, while the rest are matched with previously processed phrases or token in category 3, 4, and 5. The stream of tokens at the end of algorithm tunnel is placed in a spreadsheet for ontology analysis.

Instance Manager is used to 1) modify basic information of an entity or instance (name/supertype/description), 2) pseudo-formalize the instance description with a local pool of instance IUI codes (intra-statement), 3) determine uniqueness of an instance by searching across statement sections, and 4) clone the first appearance of an instance if current instance is not unique. This requires a manual procedure during which each screened token will be manually analyzed and annotated with 1) a unique identification code, 2) a description or definition, 3) a pseudo-formulization, 4) a supertype class (ref. to ontology database), 5) a relation to other tokens, 6) the type of relation, 7) the time when the relation obtains.

9.2 Examples

The first question/instruction of the RDC/TMD Supplemental History/Initial Questionnaire is '*In the last month, have you had any of the symptoms below? (Select ALL responses that apply for each area.)*'. The areas the patient can refer to are: Left Jaw Muscle, Left Jaw Joint, Left Ear, Left Temple, Right Jaw Muscle, Right Jaw Joint, Right Ear, and Right Temple. The symptoms that can be referred to are: Stiffness/Tightness, Cramping, Fatigue, Pressure, Soreness/Tenderness, Ache/dull ache, Throbbing, Sharp, Shooting/stabbing, Burning, Other, No symptoms in this area. This one question alone accounts for 8 areas x 12 symptoms = 96 symptom/area combinations. The outcome of asking the first of these combinations to a patient can be paraphrased as '*When the patient was asked whether at* ***feature (time, in the last month)*** *he experienced* ***feature (Symptom, Stiffness/Tightness)*** *in* ***feature (Area, Left Jaw Muscle)****, he either confirmed or not*'. Every other combination can be formed using the same pattern by substituting the ***feature (feature type, feature value)*** parts in the template with the appropriate phrases.

Table 1 shows the result of the ontological analysis performed on the first of these combinations, with the legend thereof in Table 2. While this analysis for this first combination had largely to be done manually, the analysis of the other combinations could be done largely automatically using the tools developed. Entities in green background are those entities which are explicitly referenced in the question, where those in pink are implicitly referenced.

Table 3 shows an example from the MPI for a question expecting a scaled response answer.

Table 1: ontological representation of the entities involved in answering subquestion 1 of the RDC/TMD Supplemental History/Initial Questionnaire

Ontology CLASS	Alt Class	tag 1	SuperType	UorDC	IUI-	IUICode	IDCode	Instance Origin	tag 2	INSTANCE DESCRIPTION	PSEUDO-FORMALIZATION	tag 3	Relation Domain	RelID	Relation	RelationRange	preposition	Relation Time
patient	the patient		object	DC	10	IUI-10		1.1.1	bru	The person who was being asked the question.	The person who was being asked (IUI-15).		IUI-10	r235	audience of	IUI-13	at	IUI-12
cognitive being		ie	material information	DC	11	IUI-11		1.1.1	bru	the entity presented the question to the patient.	the entity presented (IUI-15) to (IUI-10).		IUI-11	r236	actor of	IUI-13	at	IUI-12
time interval of asking a question		ie	scattered_ spatiotemporal_region	DC	12	IUI-12		1.1.1		the time interval that the patient was asked the question.	the time interval that (IUI-10) was asked (IUI-15).		IUI-12					
asking a question			study design	DC	13	IUI-13		1.1.1		the act of asking the patient the question.	the act of asking (IUI-10) (IUI-15).		IUI-13	r237	has ICE of	IUI-15	at	IUI-12
Indexicalized question			ICE	DC	14	IUI-14		1.1.1		The question 'in the last month, have you had any of the stiffness or tightness in left jaw muscle?' was presented to the patient.	The IUI-15 'IUI-21, have you had any of the IUI-25?' was presented to (IUI-10).		IUI-14					
question	the question		ICE	DC	15	IUI-15		1.1.1		> The question 'in the last month, have you had any of the stiffness or tightness in left jaw muscle?' was presented to the patient.	> The question 'in the last month, have you had any of the stiffness or tightness in left jaw muscle?' was presented to the patient.		IUI-15	r239	has indexicalized form of	IUI-14	at	IUI-16
time interval for question creation		ie	scattered_ spatiotemporal_region	DC	16	IUI-16		1.1.1		the time interval that the question were created	the time interval that (IUI-15) were created		IUI-16					
question bearer		ie	injury	DC	17	IUI-17		1.1.1		the bearer of the question on a piece of paper	the bearer of (IUI-15) on a piece of paper		IUI-17	r121	is bearer of	IUI-15	at	IUI-16
questionnaire		ie	ICE	DC	18	IUI-18		1.1.1	bru	the questionnaire that contained the questions	the IUI-18 that contained the questions		IUI-18	r142	has part	IUI-15	at	IUI-19
time interval for questionnaire creation		ie	scattered_ spatiotemporal_region	DC	19	IUI-19		1.1.1	bru	the time interval that the questionnaire was created	the time interval that the IUI-18 was created		IUI-19					
questionnaire bearer		ie	injury	DC	20	IUI-20		1.1.1	bru	the bearer of questionnaire on a piece of paper	the bearer of IUI-18 on a piece of paper		IUI-20	r121	is bearer of	IUI-18	at	IUI-19
in the last month			scattered_ spatiotemp	DC	21	IUI-21		1.1.1		the time that the patient experienced the symptom	the time that (IUI-10) experienced the symptom		IUI-21					
stiffness or tightness	x		vital signs	DC	22	IUI-22		1.1.1	bru	a physical state of being infexible and difficult to bend	a physical state of being infexible and difficult to bend	ui	IUI-22					
body area	x		constitution	DC	23	IUI-23		1.1.1	bru	an area of patient 's body	an area of IUI-10 's body	ui	IUI-23					
left jaw muscle	x		constitutional genetic	DC	24	IUI-24		1.1.1	bru	a set of muscles controlling the left side of jaw	a set of muscles controlling the left side of jaw	ui	IUI-24					
stiffness or tightness in left jaw muscle			vital signs	DC	25	IUI-25		1.1.1		a state of being inflexible in the area of left jaw muscle	a state of being inflexible in the area of left jaw muscle	ui	IUI-25	r27	develops in	IUI-24	at	IUI-21
interpreting the question		ie	bodily process	DC	26	IUI-26		1.1.1		the process that the patient understands the question is about whether the patient experienced stiffness or tightness in left jaw muscle in the last month	the process that (IUI-10) understands (IUI-15) is about whether (IUI-10) experienced IUI-25 IUI-21		IUI-26	r137	occurs in	IUI-10	at	IUI-12
assessing			bodily process	DC	27	IUI-27		1.1.1		the act of assessing about whether the patient experienced stiffness or tightness in left jaw muscle in the last month	the act of IUI-27 about whether (IUI-10) experienced IUI-25 IUI-21		IUI-27	r137	occurs in	IUI-10	at	IUI-12
answering a question			study design	DC	28	IUI-28		1.1.1	bru	The process of giving or not giving a response by the patient.	The process of giving or not giving a response by (IUI-10).		IUI-28	r74	has participant	IUI-10	at	IUI-12

Table 2: Legend for Table 1

About Entities

#	Column	Labels	Content
4	D	Ontology CLASS	Entity names
5	E	Alt Class	Alternative entity names
6	F	tag 1	inexplicit instance marker (ie)
7	G	SuperType	Supertype from SuperType Library
8	H	UorDC	Universal or Defined Class (U, DC)
9	I	IUI-	IUI number
10	J	IUICode	IUI code
11	K	ID-	ID number
12	L	IDCode	ID code
13	M	InstanceOrigin	the origin of a instance (a question number in column B)
14	N	tag 2	reused instance marker (re)
15	O	INSTANCE DESCRIPTION	Description about this instance
16	P	PSEUDO-FORMALIZATION	Pseudo-formalization of instance description
17	Q	tag 3	Customized marker

About Relations

#	Column	Labels	Content
18	R	RelationDomain	IUI-code of this instance
19	S	RelID	Relation Id from Relation Library
20	T	RelationOntology	Relation Ontology from Relation Library
21	U	RelationRange	IUI-code of the Range instance
22	V	preposition	preposition of time instance
23	W	RelationTime	IUI-code of the time instance

Table 1: ontological representation of the entities involved in answering subquestion 1

Ontology CLASS	Alt Class	tag	SuperType	UorD	IUI-	IUICode	ID-	IDCode	Instance	tag	INSTANCE DESCRIPTION	PSEUDO-FORMALIZATION	tag	RelationD	RelID	RelationOntol	Relation	pre	Relation
patient	the patient		organism	DC	7	IUI-7			1.1	bru	The person who was being asked a question	The person who was being (IUI-10)		IUI-7	r235	audience of	IUI-10	at	IUI-9
cognitive being		ie	object	DC	8	IUI-8			1.1	bru	the entity presented the question to the patient.	the entity presented the IUI-11 to (IUI-7).		IUI-8	r236	actor of	IUI-10	at	IUI-9
time Interval of asking a question		ie	temporal_region	DC	9	IUI-9			1.1		the time interval of asking a question	the time interval of IUI-10		IUI-9					
asking a question	asked a question		process	DC	10	IUI-10			1.1		the act of asking the patient a question	the act of asking (IUI-7) a IUI-11		IUI-10	r237	has ICE of	IUI-11	at	IUI-9
question			ICE	DC	11	IUI-11			1.1		> the question 'rate the level of your pain at the present moment.' was presented to the patient	> the question 'rate the level of your pain at the present moment.' was presented to the patient		IUI-11	r239	has indexicalized form of	IUI-12	at	IUI-9
Indexicalized question			material entity	DC	12	IUI-12			1.1		the question 'rate the level of your pain at the present moment.' was presented to the patient	the IUI-11 'rate the level of your IUI-19 at IUI-18.' was presented to the IUI-7		IUI-12					
question bearer		ie	material entity	DC	13	IUI-13			1.1		the bearer of the question on a piece of paper	the bearer of the IUI-11 on a piece of paper		IUI-13	r121	is bearer of	IUI-11	at	IUI-14
time interval of question creation		ie	temporal_region	DC	14	IUI-14			1.1		the time interval that the question was created	the time interval that the IUI-11 was created		IUI-14					
questionnaire		ie	ICE	U	15	IUI-15			1.1	bru	the questionnaire that contained the question	the IUI-15 that contained the IUI-11		IUI-15	r142	has part	IUI-11	at	IUI-14
questionnaire bearer		ie	material entity	DC	16	IUI-16			1.1	bru	the bearer of questionnaire on a piece of paper	the bearer of IUI-15 on a piece of paper		IUI-16	r121	is bearer of	IUI-15	at	IUI-17
time interval of questionnaire creation		ie	temporal_region	DC	17	IUI-17			1.1	bru	the time interval that the questionnaire was created	the time interval that the IUI-15 was created		IUI-17					
the present moment			temporal_region	DC	18	IUI-18			1.1		the present moment that the patient was being asked to respond	the present moment that (IUI-7) was being asked to respond		IUI-18					
pain			pain	U	19	IUI-19			1.1		The feeling of the patient's discomfort	The feeling of (IUI-7)'s discomfort		IUI-19	r137	occurs in	IUI-7	at	IUI-18
interpreting the question		ie	mental process	DC	20	IUI-20			1.1		the process that the patient understands the question is about pain	the process that (IUI-7) understands the IUI-11 is about IUI-19		IUI-20	r137	occurs in	IUI-7	at	IUI-9
assessing			mental process	DC	21	IUI-21			1.1		the act of assessing the level of pain at the present moment by the patient	the act of IUI-21 the level of IUI-19 at IUI-18 by the IUI-7		IUI-21	r137	occurs in	IUI-7	at	IUI-9
response matching		ie	appraisal process	DC	22	IUI-22			1.1		the process that the patient matches the result of assessing the pain at the the present moment to one of the items in the allowed answer list 01	the process that (IUI-7) matches the result of IUI-21 the IUI-19 at the IUI-18 to one of the items in the IUI-24		IUI-22	r137	occurs in	IUI-7	at	IUI-9
answering a question			process	DC	23	IUI-23			1.1	bru	The process of giving or not giving a response by the patient.	The process of giving or not giving a response by (IUI-7).		IUI-23	r74	has participant	IUI-7		

allowed answer list 01			data set	DC	24	IUI-24			1.1		A collection of possible answers "no pain, ...very intense pain"	A collection of possible answers "no IUI-19, ...very intense IUI-19"		IUI-24	r19	derives from	IUI-11	at	IUI-14
Response descriptor 1 - Pain Intensity		rd	data item	DC	25	IUI-25			1.1		A descriptor of pain intensity represented by 0 , from scale of 0 to 6 in allowed answer list 01	A descriptor of IUI-19 intensity represented by 0 , from scale of 0 to 6 in IUI-24		IUI-25	r160	part of	IUI-24	at	IUI-14
Response descriptor 2 - Pain Intensity			data item	DC	26	IUI-26			1.1		A descriptor of pain intensity represented by 1 , from scale of 0 to 6 in allowed answer list 01	A descriptor of IUI-19 intensity represented by 1 , from scale of 0 to 6 in IUI-24		IUI-26	r160	part of	IUI-24	at	IUI-14
Response descriptor 3 - Pain Intensity			data item	DC	27	IUI-27			1.1		A descriptor of pain intensity represented by 2 , from scale of 0 to 6 in allowed answer list 01	A descriptor of IUI-19 intensity represented by 2 , from scale of 0 to 6 in IUI-24		IUI-27	r160	part of	IUI-24	at	IUI-14
Response descriptor 4 - Pain Intensity			data item	DC	28	IUI-28			1.1		A descriptor of pain intensity represented by 3 , from scale of 0 to 6 in allowed answer list 01	A descriptor of IUI-19 intensity represented by 3 , from scale of 0 to 6 in IUI-24		IUI-28	r160	part of	IUI-24	at	IUI-14
Response descriptor 5 - Pain Intensity			data item	DC	29	IUI-29			1.1		A descriptor of pain intensity represented by 4 , from scale of 0 to 6 in allowed answer list 01	A descriptor of IUI-19 intensity represented by 4 , from scale of 0 to 6 in IUI-24		IUI-29	r160	part of	IUI-24	at	IUI-14
Response descriptor 6 - Pain Intensity			data item	DC	30	IUI-30			1.1		A descriptor of pain intensity represented by 5 , from scale of 0 to 6 in allowed answer list 01	A descriptor of IUI-19 intensity represented by 5 , from scale of 0 to 6 in IUI-24		IUI-30	r160	part of	IUI-24	at	IUI-14
Response descriptor 7 - Pain Intensity			data item	DC	31	IUI-31			1.1		A descriptor of pain intensity represented by 6 , from scale of 0 to 6 in allowed answer list 01	A descriptor of IUI-19 intensity represented by 6 , from scale of 0 to 6 in IUI-24		IUI-31	r160	part of	IUI-24	at	IUI-14
Selection level 1 - Pain Intensity		sl	directive information entity	DC			32	ID-32	1.1		When the patient was asked to rate the level of the patient's pain at the present moment, the patient answered '1', where '1' was selected from the allowed answer list 01 in scale of 1 to 6.	When (IUI-7) was asked to rate the level of (IUI-7)'s IUI-19 at IUI-18, (IUI-7) answered '1', where '1' was selected from the IUI-24 in scale of 1 to 6.		ID-32	r109	inheres in	IUI-7	at	IUI-9
Selection level 2 - Pain Intensity			directive information entity	DC			33	ID-33	1.1		When the patient was asked to rate the level of the patient's pain at the present moment, the patient answered '2', where '2' was selected from the allowed answer list 01 in scale of 1 to 6.	When (IUI-7) was asked to rate the level of (IUI-7)'s IUI-19 at IUI-18, (IUI-7) answered '2', where '2' was selected from the IUI-24 in scale of 1 to 6.		ID-33	r109	inheres in	IUI-7	at	IUI-9
Selection level 3 - Pain Intensity			directive information entity	DC			34	ID-34	1.1		When the patient was asked to rate the level of the patient's pain at the present moment, the patient answered '3', where '3' was selected from the allowed answer list 01 in	When (IUI-7) was asked to rate the level of (IUI-7)'s IUI-19 at IUI-18, (IUI-7) answered '3', where '3' was selected from the IUI-24 in scale of 1 to 6.		ID-34	r109	inheres in	IUI-7	at	IUI-9

Selection level 4 - Pain Intensity			directive information entity	DC			35	ID-35	1.1		When the patient was asked to rate the level of the patient's pain at the present moment, the patient answered '4', where '4' was selected from the allowed answer list 01 in scale of 1 to 6.	When (IUI-7) was asked to rate the level of (IUI-7)'s IUI-19 at IUI-18, (IUI-7) answered '4', where '4' was selected from the IUI-24 in scale of 1 to 6.		ID-35	r109	inheres in	IUI-7	at	IUI-9
Selection level 5 - Pain Intensity			directive information entity	DC			36	ID-36	1.1		When the patient was asked to rate the level of the patient's pain at the present moment, the patient answered '5', where '5' was selected from the allowed answer list 01 in scale of 1 to 6.	When (IUI-7) was asked to rate the level of (IUI-7)'s IUI-19 at IUI-18, (IUI-7) answered '5', where '5' was selected from the IUI-24 in scale of 1 to 6.		ID-36	r109	inheres in	IUI-7	at	IUI-9
Selection level 6 - Pain Intensity			directive information entity	DC			37	ID-37	1.1		When the patient was asked to rate the level of the patient's pain at the present moment, the patient answered '6', where '6' was selected from the allowed answer list 01 in scale of 1 to 6.	When (IUI-7) was asked to rate the level of (IUI-7)'s IUI-19 at IUI-18, (IUI-7) answered '6', where '6' was selected from the IUI-24 in scale of 1 to 6.		ID-37	r109	inheres in	IUI-7	at	IUI-9

9.3 Results

9.3.1 RDC/TMD Supplemental History: Initial and follow-up Questionnaire (RDC-SH)

The RDC/TMD (Research Diagnostic Criteria for Temporomandibular Disorders) is a dual-axis diagnostic system for Temporomandibular Disorders. It contains a well-operationalized history and examination protocol. The supplemental history initial questionnaire with 68 questions and follow-up questionnaire with 24 questions are designed to collect information from 5 domains: pain and related symptoms, headaches, jaw joint noises, jaw locking or catching closed, jaw locking or catching open. The responses to the questions can be: symptoms, area with symptoms, duration or episodicity of symptoms, frequency of symptoms, activity related to the symptoms, occurrence of symptoms, treatment or injury that may cause the symptoms.

The method described resulted in 1,062 ontologically analyzable strings from 68 questions/responses of the initial questionnaire and 24 questions/responses of the follow-up questionnaire. The analysis process identified 20,925 possible entities and 14,028 relationships between the entities.

9.3.2 Multidimensional pain inventory (Kerns, Turk, & Rudy, 1985)

The MPI questionnaire has 61 questions in 3 sections. Section 1 (28 questions) assesses pain and how it affects the patient's life (responses: 0-6, 0 = not at all and 6 = extremely). Section 2 (14 questions) assesses how the patient's spouse (or significant other) responds to the patient when he or she knows he is in pain. (responses: 0-6, 0 = never and 6 = very often) while section 3 (19 questions) assesses the performance frequency of certain daily activities (responses: 0-6, 0 = never and 6 = very often)

There are 61 ontology analyzable strings from 61 questions. Since the responses are consistently graded from 0 to 6, the responses are categorized as tokens directly. The Ontology Analysis process discovered 2,042 identifiable entities and 1,784 relationships between the entities.

9.3.3 Symptom Checklist 90R (SCL-90R)

The SCL-90R questionnaire contains 90 questions about problems in the patient's life in the past 7 days including the day of answering the questionnaire. The patient is asked to select responses that best describe how much the problem has distressed or bothered him or her as following: 0 = not at all, 1 = a little bit, 2 = moderately, 3 = quite a bit, 4 = extremely.

There are 90 ontology analyzable strings from 90 questions. Since the responses are consistently graded from 0 to 4, the responses are categorized as tokens directly. After Ontological Analysis, there are 1,712 identifiable entities and 1,259 relationships between the entities.

9.3.4 State-trait Anxiety Inventory (STAI)

The State-trait Anxiety Inventory questionnaire contains 20 questions about problems related to the patient's current state. The patient is asked to select responses that best describe the state as following: 1 = almost never, 2 = sometimes, 3 = often, 4 = almost always.

There are 20 ontology analyzable strings from 20 questions. Since the responses are consistently graded from 1 to 4, the responses are categorized as tokens directly. After Ontology Analysis, there are 526 identifiable entities and 446 relationships between the entities.

9.3.5 SF-12 Health Survey

The SF-12 Health Survey contains 7 questions and 6 contingency questions. The responses include General: 1-5, 1 = Excellent and 5 = poor, Limitation: 1-3, 1 = limited a lot and 3 = not limited at all, Time: 1-5, 1 = all the time and 5 = none of the time, Pain interference: 1-5, 1 = not at all and 5 = extremely.

There are 13 ontology analyzable strings from 14 questions. The Ontology Analysis process reveals 242 identifiable entities and 157 relationships.

9.3.6 Comparison table

		RDC-SH	MPI	SCL-90R	STAI	SF-12
	Categories	N	N	N	N	N
1	Questions	92	61	90	20	13
2	Ontology analyzable strings	1,062	61	90	20	13
3	Meaningful Tokens	20,925	2,042	1,712	526	242
	Type: Universal	1,526	116	182	53	3
	Unique/1st occurrence	**188**	**103**	**4**	**39**	**1**
	repeated	1,338	13	178	14	2
	Type: DC	19,397	1,914	1,529	472	238
	Unique/1st occurrence	**10,660**	**454**	**446**	**95**	**83**
	repeated	8,737	1,460	1,083	377	155
	Type: Response item		366	400	80	13
	Unique		3	1	1	5
	repeated		363	399	79	8
4	Relations	14,028	1,784	1,259	446	157

10 Some ontological principles for the development and curation of classifications using the International Classification of Headache Disorders as an example[2]

10.1 Ontology-Based Classification

Even when clinicians and biomedical researchers are experts in their domain, there is no guarantee that they are also experts in designing terminologies or classifications for use in their domain. That the publication of a (new version of a) classification is based on consensus is also not a guarantee for quality. Moreover, quality is usually measured in terms of (1) how well users are able to classify cases in the same way, (2) whether all cases can be classified - an easy solution to guarantee this being the introduction of 'other' or 'not elsewhere classified' type of classes - or, (3) if the classification uses criteria, whether following the criteria may nevertheless lead to cases being classifiable in more than one class such that, in case of a diagnostic classification, a patient may be diagnosed as having two disorders at the same time while there is no evidence for that being the case. Quality from an Ontological Realism perspective is more demanding. It means for classifications that the definitions for classes must follow certain principles, and that these classes correspond to pre-defined ontological categories. If the classification is designed for the medical domain, then the classes should be based on OGMS. The main goal for these extra quality criteria is to ensure that ontology-based classifications cannot only reliably be used by humans, but also that datasets collected in their terms can be fully integrated.

10.2 Recommendations

It is painful to see how currently well-known and widely used pain classifications fall short of good ontological and even terminological design in many respects. This will be illustrated by listing some important principles and demonstrating how these principles are violated in the International Classification of Headache Disorders (ICHD) (http://ihs-classification.org/en/), specifically in the newly revised Chapter 13.

10.2.1 P1: Be explicit whether assertions are about particulars or types.

Ontological Realism distinguishes between *particulars* (entities that carry identity such as me and the headache I suffered from yesterday evening) and *types* (such as human being and pain) of which the former are instances. Assertions should be construed in such a way that the terms used therein are unambiguous, including whether types or particulars are intended. The description for '*13.11 Persistent idiopathic facial pain (PIFP)*' which reads '*persistent facial pain with varying presentations and without clinical neurological deficit*' violates this principle. The term '*persistent facial pain*' in the latter can be interpreted as denoting a particular - though an arbitrary one as clearly not the specific pain of a specific patient is intended here - which means that for a specific patient to have such a pain, that pain - i.e. that very same patient's pain and not some other patient's pain - should present itself in various ways, for instance dull now, throbbing then, and so forth to qualify for being an instance of the type PIFP. But the term can also be interpreted as denoting a type in which case instances can be themselves invariant, thus some instances being dull, others throbbing, and so forth.

[2] Material presented as *Ontology: innovative approach to orofacial pain classification*. IADR Satellite Symposium on Orofacial Pain Assessment: Classification, Biobehavior, QST, and Biomarkers, March 19, 2013, Seattle, WA.

10.2.2 P2: Be precise about the sort of particulars to be classified using the classification.

The ICHD and its documentation do not present a coherent view of what the most generic type of which all particulars to be classified should be instances of might be. In the preface we are first told it is *disorders* and later *patients*, while some of the definitions indicate that it is *pains*. The recently revised Chapter 13 has as title '*Painful cranial neuropathies and other facial pains*', thus indicating that is both *pains* and *disorders* that are classified therein. Inspection of the hierarchy adds other types to the mix such as, for example, *palsies* and *syndromes*. Although certain instances of patients, pains, palsies, syndromes and disorders are related to each other, most of these instances cannot be instance of more than one of these types. It makes therefore no sense to classify all these entities in a mono-axial system.

10.2.3 P3: Particulars that correctly can be classified at a certain class level, and thus are instances of the corresponding type, should also be instance of all the types that correspond with higher level classes.

The newly revised Chapter 13 exhibits several violations of this principle. It lists for example the class '*13.1.2 Painful Trigeminal Neuropathy*' as a subclass of '*13.1. Trigeminal Neuralgia*'. While '*Neuralgia'* is defined as being pain in the distribution of nerve(s) and pain as a sensorial and emotional experience, a '*Neuropathy'* is defined as a disturbance of function or pathological change in a nerve. There is no way that one can be a special kind of the other as emotional experiences do not happen in the distribution of a nerve. Of course, when a neuropathy is painful, there is an emotional experience *involved*, i.e. *related* to the neuropathy, but that does not mean that the neuropathy *is* an emotional experience.

10.2.4 P4: Keep knowledge separate from what the knowledge is about.

Several classes have labels of the form '*X attributed to Y*', as in '*13.1.2.4 Painful Trigeminal neuropathy attributed to MS plaque*' which is then further described as '*Trigeminal neuropathy **induced** by MS plaque*' (note that '*attributed to'* is not consistent with '*induced by'*, an issue dealt with in P5). '*Attributed'* means, in this case, that it is somebody's opinion that the neuropathy is caused by MS plaque, leaving open the possibility that the neuropathy is not caused by MS plaque at all. The problem here is that a feature on the side of the clinician - his believing, probably with some degree of confidence - is presented as if it were a feature of the neuropathy, which is of course absurd. Each instance of neuropathy either is, or is not induced by MS plaque. It is true that this sort of classes are pervasive in classification systems but they nevertheless rest on a mistake: a confusion of ontology with epistemology [90].

10.2.5 P5: Class descriptions should be consistent with class labels.

There are several instances where the descriptions contain conflicting (see example in P4), inaccurate or incomplete (e.g. '*13.1.2.4 Painful Trigeminal neuropathy attributed to MS plaque*' leaves the pain out in the description) information compared to the class label. Sometimes it is additional information. It would make sense to be more consistent in the use of what is called '*description'*.

10.2.6 P6: Use Aristotelian definitions.

Classes should have - in addition to a label and a description - a definition which provides the necessary and sufficient conditions for an instance to be a member of the corresponding class. These definitions should be in Aristotelian form, roughly: an X is a Y which Z, where Y is the immediate less specific class above X. An example would be: *a Painful Post Traumatic*

Trigeminal Neuropathy is a Painful Trigeminal Neuropathy which occurs after trauma (or is caused by trauma, whatever the domain experts feel appropriate). Definitions of this form prevent odd shifts to happen such as between '*13.3.2. Secondary Nervus Intermedius Neuropathy attributed to Herpes Zoster*' and '*13.3 Nervus Intermedius (Facial Nerve) Neuralgia*' which would lead to the rather odd Aristotelian definition (shortened) 'a … Neuropathy … is a … Neuralgia … which is attributed to Herpes Zoster'" no neuropathy can *be* a pain.

10.2.7 P7: Clinical criteria do not replace Aristotelian definitions.

Whereas definitions should describe what the entities that fall under a class ***are***, clinical criteria help in ***recognizing*** whether a particular entity might fall under the class. Such criteria are typically more restrictive than definitions should be. '*13.1.1.1 Classical trigeminal neuralgia, purely paroxysmal*', for example, exhibits the criterion '*at least three attacks of facial pain fulfilling criteria B-E*'. This criterion should not be interpreted to mean that patients who had only two such attacks do not have this form of neuralgia. They might indeed have the disorder, but the criterion does not allow a clinician to make the - perhaps correct - diagnosis. This line of thinking applies to all time-related criteria, an often encountered one being the criterion for chronic pain as pain that is present for longer than three months: if a patient does suddenly have a pain for the first time in his life, it might very well be a chronic pain, but we have no way to tell at that point in time whether that is the case unless we wait three months. If so, it would also be wrong to state that the patient's pain *became* chronic after three months since, again, it was chronic all the time, but we didn't know.

10.3 Conclusions

We have outlined - without being exhaustive - a number of important ontology-based principles for building classifications. We also have shown that they are violated by the newly revised chapter 13 of the ICHD. It is easy to show that they are violated throughout the entire ICHD. Although we recognize that the ICHD in its current form is better for the advance of research than no headache classification at all, its usefulness for making patient data automatically comparable cross institutions and linguistic borders can be improved dramatically if the principles were applied.

11 Pain Assessment Terminology in the NCBO BioPortal

11.1 Introduction

Findings based on the various kinds of responses that patients may report when subjected to stimuli to test their somatosensory status, are typically described using terms such as '*allodynia'*, '*hyperesthesia'*, and so forth.

TABLE I. PAIN TERMS ANALYZED

Allodynia: pain due to a stimulus that does not normally provoke pain. Note: The stimulus leads to an unexpectedly painful response.
Analgesia: absence of pain in response to stimulation which would normally be painful.
Dysesthesia: an unpleasant abnormal sensation, whether spontaneous or evoked. Note: Special cases of dysesthesia include hyperalgesia and allodynia.
Hyperalgesia: increased pain from a stimulus that normally provokes pain.
Hyperesthesia: increased sensitivity to stimulation, excluding the special senses. Note: Hyperesthesia includes both allodynia and hyperalgesia, but the more specific terms should be used wherever they are applicable.
Hyperpathia: a painful syndrome characterized by an abnormally painful reaction to a stimulus.
Hypoalgesia: diminished pain in response to a normally painful stimulus.
Hypoesthesia: decreased sensitivity to stimulation, excluding the special senses.
Paresthesia: an abnormal sensation, whether spontaneous or evoked. Note: it has been agreed to recommend that paresthesia be used to describe an abnormal sensation that is not unpleasant while dysesthesia be used preferentially for an abnormal sensation that is considered to be unpleasant. There is a sense in which, since paresthesia refers to abnormal sensations in general, it might include dysesthesia,

The BioPortal of the National Center for Biomedical Ontology (NCBO) [91] contains to date 370 representational artifacts with over 5.6 million classes. The objectives of the work reported on here were to assess how these resources cover pain assessment terminology.

11.2 Methodology

The nine terms – henceforth called '*search terms*' – from Table 1 were submitted to the on-line version of the BioPortal Annotator [92] thereby using the following annotator options: (1) 'longest match only' unselected, (2) manual mappings included, and (3) inclusion of all ancestors. With these options thus set, the annotator returned for each search term ***ST*** in this step one or more records, each such record containing (1) the unique identifier of a class ***CL*** in relation to which ***ST*** was found (2) the name of the representational artifact ***RA*** to which ***CL*** belongs, (3) whether ***CL*** was retrieved on the basis of what the annotator qualifies as a '*direct match*' between ***ST*** on the one hand and a preferred term, synonym or identifier of ***CL*** on the other hand, or on the basis of being – mostly within ***RA***, but occasionally also within a representational artifact other than ***RA*** – an *ancestor* of a class which matches directly, and (4) the preferred term ***PT*** of ***CL*** [93].

In a second step, all detailed terminological information available for each ***CL*** matching directly was retrieved, including a visualization of the subsumption graph and all the mappings – if any at all – of ***CL*** to classes in other representational artifacts within the BioPortal. The raw data and

analysis file is available as [93]. Mappings between classes from different representational artifacts are further qualified by the BioPortal as being the result of enjoying shared Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS), and/or being automatically generated using the Lexical OWL Ontology Matcher (LOOM), which generates mappings based on lexical similarity of the preferred name and synonyms between pairs of ontologies [94].

To assess the extent to which the search terms are adequately covered in the individual BioPortal resources, and in the BioPortal as a whole, well-known quality assessment criteria and recommendations – see results and discussions for details – for terminologies [65, 95] and ontologies [96] were used. To assess the adequacy of the backbone hierarchy within individual resources 7 disjoint collections of in total 10 high level *groupings,* inspired by the various preferred terms that were retrieved, were constructed: [*Adverse event*], [*Body part*], [*Discipline*], [*Disease, Disorder or Finding*; *NON-pain disorder*; *Pain / sensation finding*], [*Pharm. Effect / Endpoint*], [*Function / Process*; *Technique / Therapy*] and [*Meta / Top*]. Each class (with disambiguation where required as for instance for 'analgesia') was classified into one of these groupings on the basis of its preferred term. Examples of classes labelled *Meta* are classes with preferred terms such as *Inactive Concept* and *Unclassified*, whereas the *Top* labelling include classes such as *Snomed CT Concept* and *Topical descriptor* [93].

The adequacy of the mappings between directly matched classes was assessed semi-automatically. Mapping records in which the semantics of at least one of the classes could not be determined, were excluded. Records where only one of the classes was marked as being *Meta*, were automatically tagged as obsolete. Records for which the preferred names of both classes were identical, except in the case of 'analgesia' given its homonymous semantics, were automatically assigned as being correct. All other cases were assessed manually.

11.3 Results

Querying for the 9 search terms in the BioPortal Annotator exactly as displayed in Table 1 returned 762 annotation records of which 113 were about in total 104 candidate annotation classes labelled by the Annotator as 'direct' and which originated from 27 different sources [93] out of the 371 total artifacts at the time this work was performed. 17 annotation records revealed that in the ICPC2, RH-MeSH and SNOMED CT some of the search terms matched directly to more than one class (Table 3, AP5 in Table 2) – thus reflecting homonymy, while 9 records showed that some of the classes were mapped to by distinct search terms (AP3 in Table 2) – thus reflecting synonymy for the terms involved within the context of that source. Ignoring capitalization, the 104 direct annotation classes exhibited in total 25 distinct preferred terms. In Table 3 it is displayed how these preferred terms are related to the original search terms in each resource.

225 additional candidate annotation records [93] were retrieved by querying for three of the spelling variants suggested by some of the retrieved preferred terms obtained by querying for the original search terms (Table 3): 77 for *hyperaesthesia*, 76 for *hypesthesia*, and 72 for *hypoaesthesia*. These records reveal that these terms match directly with 14 classes that were not matched with the original search terms, thereby bringing ICD10 on board as extra representational artifact. These records are not included in any further analysis. 649 annotation records were labelled by the Annotator as containing hierarchical ancestors of the classes matched directly, totaling 206 distinct ancestor classes with together 169 distinct preferred terms [93]. One class, labelled '*UMLS:OrphanClass*' appeared in 40 records involving the 8 representational artifacts labeled ICPC2, MESH, NDFRT, OMIM, PDQ, RCD, SNMI, and

SNOMEDCT. 1036 mapping records were retrieved for all 104 classes matched directly to the search terms, of which 71 duplicates, yielding 965 records further analyzed [93]. 399 of those records required manual assessment.

TABLE II. SUMMARY ASSESSMENT OF TERMINOLOGICAL AND ONTOLOGICAL QUALITY OF THE SEARCH TERM RELATED BIOPORTAL CLASSES RETRIEVED

Representational Artifact →		BDO	COSTART	CRISP	CSSO	CTCAE	DOID	GALEN	HP	ICD10CM	ICNP	ICPC2P	MEDDRA	MESH	MP	NCIT	NDFRT	NIFSTD	OMIM	PDQ	PHARE	RCD	RH-MESH	SNMI	SNOMEDCT	SOPHARM	SYMP	WHO-ART	Totals
Assessment Parameter	**Norm**																												
AP1 IASP search terms covered	9	1	6	2	1	2	1	0	5	1	2	2	8	3	4	4	4	1	1	0	1	5	3	7	9	4	6	2	
AP2 Number of direct class matches	>8	1	6	2	1	2	1	1	5	1	2	8	9	7	4	5	5	1	1	1	1	5	11	7	14	4	6	2	113
AP3 Direct classes with wrong IASP synonymy	0								1					2										1	3			2	9
AP4 Direct classes with definitions	=AP2			2	1	2	1		4					7	4	5	3	1									6		36
AP5 Number of direct classes with inappropriate homonymy	0											6											6		5				17
AP6 Number of additional direct classes through spelling variants	0	1											3			1						2	3						13
AP7 Number of class matches	>AP2	7	24	12	2	8	7	9	39	5	16	16	15	72	31	35	40	3	5	3	6	56	60	38	164	39	24	26	762
AP8 Foreign classes in hierarchy	0											8		7			5		1	1		5		7	6				40
AP9 Number of hierarchy classes with disjointness violations	0		7	4		2	4	1				6		15			11	1	3		4	5	2		49			24	60
Evaluation																													
Maximum number of norm violations		6	8	8	6	8	6	1	8	6	8	8	8	8	8	8	8	6	6	1	6	8	8	8	8	8	8	8	
Number of norm violations (except P8)		4	4	3	2	3	3	1	4	3	3	6	3	5	2	3	5	3	5	1	4	6	5	5	5	3	2	5	

11.3.1 Quality of BioPortal Resources Retrieved

Table 2 provides – with the exception of assessment parameter AP8 – a summary assessment of the terminological and ontological quality of the classes (and by extension of the resources from which they originate) that were retrieved for the 9 search terms. Further details about certain aspects are available in Table 3 and Table 4. 9 APs are considered, and for each AP a norm is determined. Table 2 thus illustrates that:

- only SNOMED CT covers the 9 search terms in the lexical form provided by the IASP (AP1), while MeDDRA has complete coverage if lexical variants are taken into account (AP6), (it was not checked whether resources contained atomic terms that through post-coordination would allow to express the terms),
- 5 resources do not make the distinctions in terminology made by the IASP (AP3, details in Table 3),
- 11 resources provide textual definitions for at least some of the classes (AP2, AP4),
- 3 resources exhibit inappropriate homonymy for some of the search terms (AP5),
- more than half of the resources exhibit for at least some of the search terms a hierarchy which on the basis of the face value of the preferred terms is composed of disjoint classes (AP9, details in Table 4),

- none of the representational artifacts cover the domain delineated by the IASP search terms adequately when taking all assessment parameters into account.

11.3.2 Adequacy of the NCBO BioPortal

Out of the 27 representational artifacts which have at least one class with a direct match to a search term, 22 have classes which by the BioPortal are mapped to at least one other class from another artifact. 618 of these mappings are within these 22 sources whereas 347 mappings are towards classes from 18 target representational artifacts outside these sources. Of these 18, MeDDRA and RH-MeSH are the only two that have classes directly matched with the search terms, thus reflecting the BioPortal documentation that mappings are not always bidirectional.

Table 5 quantifies the appropriateness of the mappings on the basis of our methodology. The 'B' and 'T' following the resource names in Table 5 indicate whether the resource exhibits mappings bi-directionally resp. only incoming. B-mappings are only counted once in the totals. Mappings are qualified as being excluded ('Excl.') from the analysis because of either ambiguity or missing information on the side of the classes mapped to ('T?') or being in the realm of the 22 source classifications ('S?'). 'Correct' mappings result from (1) the automatic assignment of the adequacy assessment for pairs of source and target classes with identical non-ambiguous preferred terms ('SAME'), and the manual verification of (2) classes with synonymous preferred terms, i.e. lexical variants or descriptions ('VARIANT') and (3) classes with ambiguous preferred terms. Erroneous mappings ('ERROR') are brought about by (1) automatic determination of mapping to or from inactive classes ('OBSO') and manual verification of (2a) mapping to or from classes with ambiguous meaning ('HOMONYM'), and (2b) inappropriate mappings between classes with unambiguous meanings ('WRONG'). Table 6 provides insight in the accuracy of the methods applied in the BioPortal to create mappings, i.e. whether on the basis of the UMLS Concept Unique Identifiers ('cui'), the LOOM algorithm ('loom') or both.

11.4 Discussion

During 'The Consensus Workshop: Convergence on an Orofacial Pain Taxonomy', held March 30 – April 1, 2009, Miami, Florida, which was attended by representatives from all major pain institutions, it was concluded that an adequate treatment of the ontology of pain together with an appropriate terminology, is mandatory to advance the state of the art in diagnosis, treatment and prevention [30].

As a first step, it was proposed to study the terminology and ontology of pain as currently defined. The ontological aspects have since then been covered in [97], and the underlying principles thereof been applied, for instance, in the definition of new pain-related disease entities and classifications [3, 59].

The analysis performed here is another response to the workshop's recommendations with the goal to obtain more insight in how pain assessment terminology is dealt with in representational artifacts such as widely used classification systems, terminologies, and ontologies. At the same time, it provided an opportunity to assess the usability of the NCBO BioPortal for a task of this nature, and the appropriateness of the principles and methods applied in the BioPortal to present a unified, highly standardized and ontology-like view on resources which are qua structure and underlying design principles very different.

TABLE III. MAPPING OF SEARCH TERMS TO PREFERRED TERMS IN THE REPRESENTATIONAL ARTIFACTS

Representational Artifact → / Search Term / Preferred Term	BDO	COSTART	CRISP	CSSO	CTCAE	DOID	GALEN	HP	ICD10CM	ICNP	ICPC2P	MEDDRA	MESH	MP	NCIT	NDFRT	NIFSTD	OMIM	PDQ	PHARE	RCD	RH-MESH	SNMI	SNOMEDCT	SOPHARM	SYMP	WHO-ART	Grand Total	Occurrence
Allodynia								**1**		**1**		**1**	**1**	**1**							**1**			**1**	**1**	**1**		**9**	**9**
Allodynia								1		1		1		1							1			1	1	1		**8**	**8**
Hyperalgesia													1															**1**	**1**
Analgesia		**1**	**1**			**1**	**1**					**1**	**1**	**1**	**1**	**1**			**1**	**1**	**1**	**2**	**1**	**1**	**1**			**17**	**16**
Analgesia			1				1					1	1	1						1		2	1		1			**10**	**9**
No sensitivity to pain																					1			1				**2**	**2**
pain agnosia						1																						**1**	**1**
Analgesia [PE]																1												**1**	**1**
Hypalgesia		1																										**1**	**1**
Pain Therapy															1				1									**2**	**2**
Dysesthesia					**1**			**1**				**1**	**1**		**1**								**1**	**2**				**8**	**7**
Dysesthesia					1			1				1			1								1	2				**7**	**6**
Paresthesia													1															**1**	**1**
Hyperalgesia		**1**	**1**							**1**		**1**	**1**	**1**	**1**	**1**	**1**				**1**	**3**	**1**	**3**	**1**	**1**	**1**	**20**	**16**
O/E - hyperesthesia present (& [hyperalgesia])																								2				**2**	**1**
Hyperalgesia [Disease/Finding]																1												**1**	**1**
Hyperalgesia		1	1							1		1	1	1	1		1				1	3	1	1	1	1		**16**	**14**
HYPERAESTHESIA																											1	**1**	**1**
Hyperesthesia		**1**						**1**	**1**		**4**	**1**	**1**			**1**						**3**	**1**	**1**		**1**		**16**	**11**
Hyperesthesia [Disease/Finding]																1												**1**	**1**
HYPERAESTHESIA											4																	**4**	**1**
Hyperesthesia		1						1	1			1	1									3	1	1		1		**11**	**9**
Hyperpathia												**1**									**1**		**1**	**2**				**5**	**4**
Hyperalgesia																							1	1				**2**	**2**
Hyperpathia												1									1			1				**3**	**3**
Hypoalgesia		**1**										**1**		**1**							**1**			**1**	**1**	**1**	**1**	**8**	**8**
HYPALGESIA		1																										**1**	**1**
HYPOAESTHESIA																											1	**1**	**1**
Hypoalgesia												1		1							1			1	1	1		**6**	**6**
Hypoesthesia		**1**						**1**				**1**	**1**		**1**	**1**							**1**	**2**		**1**		**10**	**9**
Hypesthesia [Disease/Finding]																1												**1**	**1**
Reduced sensation of skin																								1				**1**	**1**
Sensory impairment								1																				**1**	**1**
Hypesthesia		1											1											1				**3**	**3**
Hypoesthesia												1			1								1			1		**4**	**4**
Paresthesia	**1**	**1**		**1**	**1**			**1**			**4**	**1**	**1**		**1**	**1**		**1**				**3**	**1**	**1**		**1**		**20**	**15**
Paresthesia [Disease/Finding]																1												**1**	**1**
paraesthesia											4																	**4**	**1**
Paresthesia	1	1		1	1			1				1	1		1		1					3	1	1		1		**15**	**13**
Grand Total	**1**	**6**	**2**	**1**	**2**	**1**	**1**	**5**	**1**	**2**	**8**	**9**	**7**	**4**	**5**	**5**	**1**	**1**	**1**	**1**	**5**	**11**	**7**	**14**	**4**	**6**	**2**	**113**	
Occurrence	**1**	**6**	**2**	**1**	**2**	**1**	**1**	**5**	**1**	**2**	**2**	**9**	**7**	**4**	**5**	**5**	**1**	**1**	**1**	**1**	**5**	**4**	**7**	**9**	**4**	**6**	**2**		**95**

TABLE IV. GROUPING OF THE SEARCH TERMS IN DISJOINT UPPER CLASSES IN THE HIERARCHY OF THE REPRESENTATIONAL ARTIFACTS

Representational Artifact → / Search Term Grouping	BDO	COSTART	CRISP	CSSO	CTCAE	DOID	GALEN	HP	ICD10CM	ICNP	ICPC2P	MEDDRA	MESH	MP	NCIT	NDFRT	NIFSTD	OMIM	PDQ	PHARE	RCD	RH-MESH	SNMI	SNOMEDCT	SOPHARM	SYMP	WHO-ART	Grand Total
Allodynia								**7**		**8**		**4**	**11**	**8**							**11**			**15**	**10**	**5**		**79**
Disease or Finding								6		4		4	8	8							9			11	10	5		65
Function / Process										4																		4
Meta / Top								1					3								2			4				10
Analgesia		**4**	**5**			**7**	**9**					**1**	**6**	**7**	**6**	**8**			**3**	**6**	**11**	**5**	**6**	**9**	**9**			**102**
Body part		1																										1
Discipline																				3								3
Disease or Finding		3	2			3	1					1	1	7						2	9	3	3	9	9			53
NON-pain disorder						4																						4
Technique / Therapy			3										1		3				2									9
Function / Process							6								3								2					11
Meta / Top							2						3			2			1		2		1					11
Pharm. Eff./Endpoint													1			6				1		2						10
Dysesthesia					**4**			**9**				**1**	**11**		**8**								**5**	**22**				**60**
Adverse event					1																							1
Disease or Finding					2			8				1	8		8								2	13				42
Function / Process																							2					2
Meta / Top					1			1					3										1	9				15
Hyperalgesia		**4**	**7**							**8**		**1**	**11**	**8**	**5**	**8**	**3**				**14**	**19**	**6**	**29**	**10**	**4**	**13**	**150**
Body part		1																										1
Discipline			4																									4
Disease or Finding		3	3							4		1	8	8	5	6	1				12	19	3	15	10	4	1	103
NON-pain disorder																											12	12
Function / Process										4													2					6
Meta / Top													3			2	2				2		1	14				24
Hyperesthesia		**4**						**8**	**5**		**8**	**1**	**11**			**8**						**18**	**5**	**14**		**4**		**86**
Body part		1																										1
Disease or Finding		3						7	4		4	1	8			6						18	2	10		4		67
Function / Process																							2					2
Meta / Top								1	1		4		3			2							1	4				16
Hyperpathia												**4**									**10**		**6**	**23**				**43**
Disease or Finding												4									8		3	14				29
Function / Process																							2					2
Meta / Top																					2		1	9				12
Hypoalgesia		**4**										**1**		**8**							**10**			**17**	**10**	**4**	**13**	**67**
Body part		1																										1
Disease or Finding		3										1		8							8			13	10	4	1	48
NON-pain disorder																											12	12
Meta / Top																					2			4				6
Hypoesthesia		**4**						**7**				**1**	**11**		**8**	**8**							**5**	**20**		**4**		**68**
Body part		1																										1
Disease or Finding		3						6				1	8		8	6							2	11		4		49
Function / Process																							2					2
Meta / Top								1					3			2							1	9				16
Paresthesia	**7**	**4**		**2**	**4**			**8**			**8**	**1**	**11**		**8**	**8**		**5**				**18**	**5**	**15**		**3**		**107**
Adverse event					1																							1
Body part		2																2										4
Discipline																		1										1
Disease or Finding	6	2		2	2			7			4	1	8		8	6		1				18	2	11		3		81
Function / Process																							2					2
Meta / Top	1				1			1			4		3			2		1					1	4				18
Grand Total	**7**	**24**	**12**	**2**	**8**	**7**	**9**	**39**	**5**	**16**	**16**	**15**	**72**	**31**	**35**	**40**	**3**	**5**	**3**	**6**	**56**	**60**	**38**	**164**	**39**	**24**	**26**	**762**

TABLE V. CORRECTNES OF DIRECT CLASS MAPPINGS

Representational Artifacts		Error			Correct			Excl.		TOTAL	% WRONG
		WRONG	OBSO	HOMONYM	SAME	VARIANT	DISAMBIG.	S?	T?		
ACGT-MO	T	1								1	100
AI-RHEUM	T	1								1	100
BDO	B		3		20	5			3	31	10.7
COSTART	B	31	12		44	40			44	171	34.4
CRISP	B	7	4		17	2		19	21	70	36.7
CSSO	B	1	2		17	8			3	31	10.7
CTCAE	B		4		19	4			4	31	14.8
GALEN	B	5		9	1	7	6	1	3	32	50
HIMC-ICD09	T		2						9	11	100
HIMC-LOINC	T								3	3	0
HL7	T					4		2		6	0
HOM-CLINIC	T								3	3	0
HOMERUN-UHC	T								3	3	0
HP	B	7	2		19	2			4	34	30
ICD10	T		1		2	3				6	16.7
ICD10CM	B		2		7	6			4	19	13.3
ICPC2P	B	1	6		21	48		1	10	87	9.21
IFAR	T					2				2	0
LOINC	T					6		3		9	0
MEDDRA	T		15						122	137	100
MESH	B	2	8	5	44	13	4	1	16	93	19.7
MP	B	6	6	13	29	8	6	1	7	76	36.8
NCIT	B	9	9	4	43	23	3	1	23	115	24.2
NCIt-Activity	T	2			2					4	50
NDFRT	B	5	10	8		60	1	1	28	113	27.4
NDF-RT	T	2	4	7	24	8	1	1	2	49	37
NIFSTD	B	2	2		17	4			2	27	16
OMIM	B	3	3		18	9			2	35	18.2
PDQ	B	3		4	4	8	3	1	11	34	31.8
PHARE	B	2		11			8	1	3	25	61.9
PMA	T					2		1		3	0
RCD	B	9	10	5	36	24	4	1	11	100	27.3
RH-MESH	T		12						86	98	100
RPO	T	2				2				4	50
SNOMEDCT	B	6	150	3	2	15	4	1	4	185	88.3
SOPHARM	B	6	8	13	29	4	6	1	10	77	40.9
SYMP	B	7	14		55	17			18	111	22.6
SYN	T	2			2					4	50
TRAK	T					2		1		3	0
WHO-ART	B	34	7		10	18			17	86	59.4
Grand Total		**78**	**148**	**41**	**241**	**177**	**23**	**19**	**238**	**965**	**37.7**
% of mappings			**27.67**			**45.70**			**26.63**		

11.4.1 Are Resources in the BioPortal intrinsically flawed

As can be inferred from Table 2 and Table 3, all retrieved resources, with – at first sight – the exception of MeDDRA and SNOMED CT, seem to perform quite poorly in terms of coverage of the domain. Of course, some resources might have been designed with a specific purpose in mind and pain assessment terminology therefor being out of their scope. It is however hard to imagine for what sort of purpose a term such as *paresthesia* might be relevant and *dysesthesia* not: if one is present, all should be present. An exception is *analgesia* in the sense of a procedure rather than of a symptom: there would indeed be no place for any of the other terms in procedure terminologies. Although there are indeed a few resources retrieved for which *analgesia* is the only term matched, these resources are not restricted to procedures. Some resources turn out to exhibit a better coverage when spelling variants are used in the queries, but not to the extent that it can explain the overall lack of coverage.

Some resources, such as COSTART, MeSH and WHO-ART, suffer from the lack of discrimination between terms in pairs such as hypoalgesia/hypesthesia, hyperalgesia/hyper-esthesia, dysesthesia/paresthesia and analgesia/hypoalgesia. This was also found in SNOMED CT but only for classes that were labelled 'inactive' thus reflecting that these mistakes made in earlier versions were corrected afterwards.

15 resources exhibit through the eyes of the BioPortal a backbone structure which at least can be frowned upon (Table 4). How can *analgesia* be a kind of *nervous system* (COSTART), *communication disorder* (DOID - Human Disease Ontology), or *pharmacogenomics* (PHARE)? How can *paresthesia* be a kind of *peripheral nervous system* (OMIM), *hyperalgesia* a kind of *adrenal adenoma* (WHO-ART) or *neuroscience* (CRISP)? One can assume sloppy design on the side of the authors of these resources, or violation of the principle that preferred terms should have face validity [65]: thus in COSTART 'nervous system' might not mean *nervous system*, but rather *symptom related to the nervous system*. Or, and this leads to the next section, perhaps the BioPortal represents the structure of these resources erroneously?

11.4.2 Is the BioPortal itself, or are some design or quality assurrance principles behind it, intrinsically flawed?

That something wasn't right with the representation of WHO-ART in the BioPortal was noted by Ruttenberg in 2011 and as such acknowledged by BioPortal staff who traced the issue down to be caused by the WHO-ART source codes, but nevertheless decided nothing to do about it at that time [98]. And apparently never since: the version of WHO-ART that showed up in the work reported about in this paper was version '2013AB' which was uploaded to the BioPortal, according to the summary page, February 18, 2014, indeed without any attention to the known issues. The data presented here demonstrate further that it is not just WHO-ART of which the representation in the BioPortal is problematic with respect to the semantics of the subclass relationship, but also 14 other resources that were retrieved on the basis of the search terms (Table 2, AP9).

Another indication that the BioPortal could benefit from some quality assurance introspection comes from the finding that for 8 of the 27 resources retrieved the Annotator returned 'UMLS:OrphanClass' as ancestor for 40 of the classes matched directly (Table 2, AP8).

Also the mapping results provide serious evidence in the direction that quality improvement is required.

TABLE VI. MAPPING SOURCES

Result	cui	cui, loom	loom	Grand Total
Error	29	20	218	**267**
WRONG	5	4	69	**78**
OBSO	24	16	108	**148**
HOMONYM			41	**41**
Correct	50	23	368	**441**
SAME	2	9	230	**241**
VARIANT	48	14	115	**177**
DISAMBIG.			23	**23**
Excluded	22	17	218	**257**
S?			31	**31**
T?	22	17	187	**226**
Grand Total	**101**	**60**	**804**	**965**
% Wrong	**36.71**	**46.51**	**37.20**	**37.71**

First there is the observation that through the mappings, 16 additional resources were discovered that contain classes which map directly to classes which were retrieved by means of the search terms. This can in part be explained by the absence of the search terms in the synonym set of these additional classes, but upon further inspection, it turns out that in case of in total 255 mappings for RH-MeSH and MeDDRA, as well as for (possible) resources which according to the syntax of the URIs of the classes mapped to might be named 'HOMERUN-UHC', 'HOM-CLINIC', 'HIMC-LOINC' and 'HIMC-ICD09', the URIs returned by the annotator do not resolve at all [93]. The former 4 resources are also not listed on the BioPortal webpage as being resources it contains, yet classes from them show up in the mapping results. In case of SNOMED CT, mappings are primarily involving classes which are marked as 'inactive'.

A second observation is that – after excluding these 255 mappings as well as two others for which the meaning of the source class could not be disambiguated – still almost 38% of the mappings are inaccurate. There is no significant difference in accuracy between mappings produced using LOOM or UMLS CUIs alone. However, when both the LOOM and CUI-methods suggest a mapping, the error rate increases to over 46%, thus almost the equivalent of flipping a coin.

11.5 Limitations

The work reported on here bears certain limitations. Although the data demonstrate (1) that the domain of pain assessment terminology is poorly covered in the BioPortal resources, (2) that the way in which the BioPortal organizes the retrieved classes hierarchically using the subclass relation is debatable, and (3) that the techniques used to map these classes between resources are not quite adequate, no generalizations can be made to other domains. A further limitation is that the data were retrieved using the BioPortal website rather than the REST services. Perhaps these services offer better ways to filter inadequate data, but if that were the case, one could wonder why such filters are not used on the website.

Assessment of the correctness of the suggested hierarchy and the mappings was carried out with the quality criteria of the OBO Foundry and adherence to the principles of Ontological

Realism in mind, neither of which are universally accepted [2] yet gaining considerable attraction [99]. Thus it is quite conceivable that reviewers outside the Foundry would report lower error rates, for instance by finding it perfectly acceptable that the 'concept' of analgesia as a pharmaceutical effect in some drug is considered equivalent to the 'concept' of analgesia as a procedure performed by an anesthesiologist or as a state of a patient brought about by such procedure. At the other hand, since the review here was based by first flagging results that for sure require manual evaluation (see methodology) it might very well be that certain mapping- or ancestor records were erroneously not flagged. In that sense, the error rates presented here could very well be – modulo mistakes made by sloppiness of the reviewer – the best case scenario. Another limitation is that this study does point out the kind of mistakes and how to find them semi-automatically, but is not conclusive on whether the root cause is in the source systems, the BioPortal, or a combination of both.

11.6 Conclusion and Recommendations

Without doubt, studies such as this one could not be carried out without a resource such as the BioPortal, or would require a lot more time and effort. Evenly without doubt, the BioPortal made it possible to reach the objectives of this study which were to find out (1) whether the sources in the BioPortal provide a more adequate view on pain assessment terminology – the answer being ***no***, and (2) to what extent the BioPortal itself is a useful instrument in determining whether (1) is indeed the case – the answer being ***yes***. As a side effect, this study raises serious questions about the quality assurance principles employed in the design and management of the BioPortal, more specifically (1) about the quality of the resources the BioPortal accepts for inclusion – it might seem unfair to criticize a lack of clear best practice policies in the investigated resources while not distinguishing their different semantic expressivity, the point being however that the BioPortal itself does not allow for such distinctions and 'promotes' all resources as ontologies, (2) the suitability of representing the hierarchy of these resources by means of the subclass relation, and (3) about certain house-keeping operations. Quality seems thus far not to have been much of a concern to the BioPortal scientific community, as witnessed by the presence of only one paper in Pubmed that addresses the topic [100]. Furthermore, although the BioPortal does indeed offer a mechanism to users to make notes on the quality of BioPortal content [91], it doesn't seem to be used much: the BioPortal homepage displays a list of the 5 last notes submitted, of which the last three were submitted 7 months prior to writing this paper, all three about a 'request' issued by user *rboden* – noted in the name of 'Jesus' as contact person – to add the following new term '*We need someone with qualifications*'. It is a bad sign that spam of this kind, whether unnoticed or noted but not acted upon, is accepted.

For the BioPortal to become an instrument which is useful for other purposes than determining that its content is of poor quality the following suggestions are in order: (1) do not accept resources that violate standard subsumption principles, (2) display for each resource quality metrics, rather than mere quantity metrics, for instance the extent to which they follow the principles of ontological realism or the OBO Foundry, and (3) provide better documentation about the methods and algorithms used to present hierarchies and mappings, and about the internal quality assurance principles.

12 An alternative terminology for pain assessment

The various kinds of responses that patients may report when subjected to stimuli to test their somatosensory status are typically described using terms such as 'allodynia', 'hyperesthesia', and so forth. Although these terms were already in practice since at least the early 19th century [101], standard definitions for these terms were first proposed in 1979 [102] and are since then regularly updated by the International Association for the Study of Pain (IASP), in print for the last time in 1994 [103], with more regular electronic updates on the IASP webpage [49] the last one in May 2012 (Table 1). These definitions are further based on the IASP definition for '*pain'* as '*an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage*'.

For terms to be eligible as representational units in a realism-based ontology such as OPMQoL, they must not only (1) denote entities that can be classified following the principles of Ontological Realism [82], but also (2) be defined using Aristotelian definitions which specify the necessary and sufficient conditions for class membership, and further lead to a taxonomy based on single inheritance [104]. The goal of the work reported on here was to assess the adherence of the IASP pain assessment definitions to this second condition and to find ways for remediation if non-compliance was found.

Allodynia: pain due to a stimulus that does not normally provoke pain. Note: The stimulus leads to an unexpectedly painful response.
Analgesia: absence of pain in response to stimulation which would normally be painful.
Dysesthesia: an unpleasant abnormal sensation, whether spontaneous or evoked. Note: Special cases of dysesthesia include hyperalgesia and allodynia.
Hyperalgesia: increased pain from a stimulus that normally provokes pain.
Hyperesthesia: increased sensitivity to stimulation, excluding the special senses.
Hyperpathia: a painful syndrome characterized by an abnormally painful reaction to a stimulus.
Hypoalgesia: diminished pain in response to a normally painful stimulus.
Hypoesthesia: decreased sensitivity to stimulation, excluding the special senses.
Paresthesia: an abnormal sensation, whether spontaneous or evoked. Note: paresthesia is to be used to describe an abnormal sensation that is not unpleasant.

Table 1 - Pain terms analyzed

12.1 Methods

Based on the definitions of the terms studied – note that table 1 contains only part of the relevant notes and that the reader should for complete understanding of the analysis method consult reference [49] - an analysis framework was designed by introducing nine hierarchically organized variables reflecting the type of stimulus, the presence or absence of a response, and the type of response when present, when a patient is subjected to a pain assessment investigation. The allowed values for these variables were defined, depending on what the variable stands for, either on a nominal or ordinal scale (Table 2).

The next step consisted of identifying and representing all theoretically possible stimulus/response combinations, a part of which is displayed in Table 3.

Variable	Values
Stimulus application	Y(es)
modus M level Threshold	B(elow), O(n), A(bove)
Pain level Threshold	B(elow), O(n), A(bove)
Response to stimulus	Y(es), N(o)
modus M Response	Y(es), N(o)
modus M Intensity	L(ess), C(oncordant), H(igh)
Unpleasant response	Y(es), N(o)
Pain Response	Y(es), N(o)
Pain Intensity	L(ess), C(oncordant), H(igh)

Table 2 - Basic analysis framework variables, values and definitions

Although the maximal theoretical number of possible combinations would be 1296 (1*3*3*2*2*3*2*2*3), the actual number is only 130 because of the hierarchical organization of the variables which implements the following dependencies typical for somatosensory and pain assessment studies [105]:

1. each stimulus, whether to test either somatosensory status (e.g. temperature, pressure, pin prick, and so forth, henceforth called 'modus M') or pain sensitivity, falls under one of three disjoint categories: (1) below threshold, (2) on threshold, or (3) above threshold;
2. modus M and pain stimuli may be given selectively or together, thus resulting in 4 stimulation modes: (1) sub-threshold (for both pain and modus M), (2-3) modus M- or pain-selective, and (4) bimodal (i.e. on or supra-threshold for both modus M and pain);
3. if there is no response to a stimulus, then there are no values for the intensity of modus M sensation and pain;
4. if a response is present, it may be either (4a) selective, i.e. exclusively being unpleasant, painful, or of modus M in isolation, or (4b) combining either a modus M and non-painful unpleasant response, or a modus M and painful response;
5. all pain responses are unpleasant, thus following the IASP definition for '*pain'* as '*an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage*', but an unpleasant response does not need to be painful.

As a third step, each combination was assessed for whether it could figure as an exemplar for each of the terms of Table 1. Table 4 provides an example of this step for the IASP-definition of 'allodynia' without taking the note into account. A complication at this phase was that the definitions and notes left certain questions with respect to inclusion and exclusion criteria unanswered. It was thus for many definitions required to find meaningful subgroups and for some of these subgroups the IASP documentation did not provide enough information to assess whether they represent intended interpretations, although from a terminological and ontological perspective perfectly plausible. Table 5 shows the subgroups identified as well as the counts of stimulus/response combinations that fall under them. When subgroups were defined, the count for the (direct or indirect) parent terms were obtained by applying a Boolean OR operation on the combinations (and not the mere addition as subgroups are not necessarily mutually exclusive). This information was in a fourth step used to compute the exact overlap between these terms in function of positive and negative co-occurrence.

S	**Stimulus given**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
MT	**Modus M threshold**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
PT	**Pain Threshold**	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
R	**Response**	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
MR	**Modus M response**	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
MI	**Modus M response Intensity**	-	-	-	-	-	L	L	L	L	L	C	C	C	C	C	H	H	H	H	H
U	**Unpleasant response**	N	Y	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y
PR	**Pain Response**	-	N	Y	Y	Y	-	N	Y	Y	Y	-	N	Y	Y	Y	-	N	Y	Y	Y
PI	**Pain Response Intensity**	-	-	L	C	H	-	-	L	C	H	-	-	L	C	H	-	-	L	C	H

Table 3 - Different stimulus/response combinations possible for bimodal above (but not 'on') threshold stimulation. Legend for values: Y = Yes, N = No, B = Below threshold stimulus, O = On threshold stimulus, A = Above threshold stimulus, H = Higher than expected response intensity, C = response intensity Concordant with stimulus, L = Lower than expected response intensity.

S	**Stimulus given**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
MT	**Modus M threshold**	B	B	O	O	O	O	A	A	A	A
PT	**Pain Threshold**	B	B	B	B	B	B	B	B	B	B
R	**Response**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
MR	**Modus M response**	N	Y	N	Y	Y	Y	N	Y	Y	Y
MI	**Modus M response Intensity**	-	H	-	L	C	H	-	L	C	H
U	**Unpleasant response**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PR	**Pain Response**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
PI	**Pain Response Intensity**	H	H	H	H	H	H	H	H	H	H
ALLO-D		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Table 4 - Possible stimulus/response combinations for Allodynia (following the IASP definition strictly). Legend for values: Y = Yes, N = No, B = Below threshold stimulus, O = On threshold stimulus, A = Above threshold stimulus, H = Higher than expected response intensity, C = response intensity Concordant with stimulus, L = Lower than expected response intensity.

This step answers thus for each term pair 'A B' the question which and how many of the possible stimulus/response combinations can occur in the pair combinations A+/B+, A+/B-, A-/B+,A-/B- where '+' and '-' indicate that the stimulus/response combination can, resp. cannot occur under the definition of the term. As it became clear at this point that overlap was considerable, we designed a new terminology based on definitions that minimize the potential overlap using categories that are mutually exclusive. We then compared this new terminology with the traditional one, again using the stimulus/response combinations as benchmark.

12.2 Results

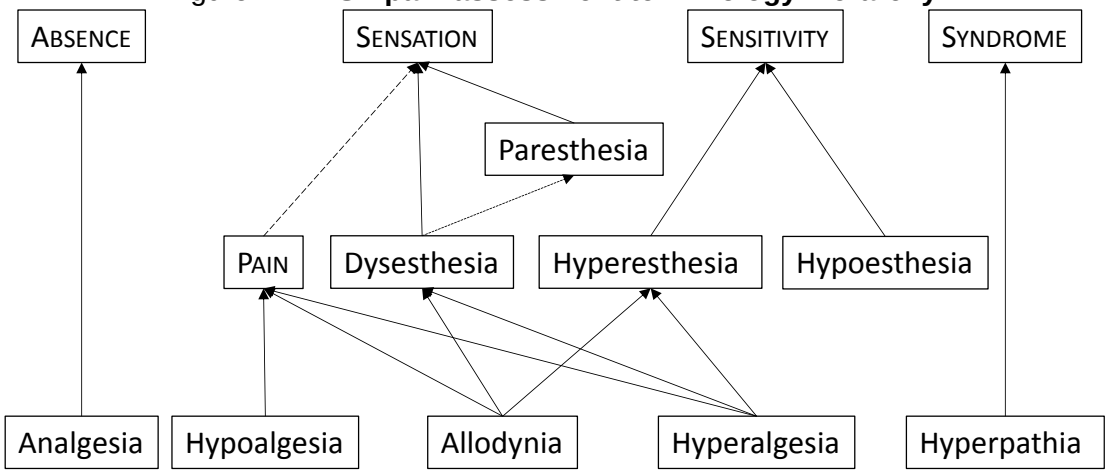
12.2.1 The IASP terms do not satisfy the criteria for direct integration in a realism-based ontology.

Figure 1 - in which terms displayed in SMALL CAPS are the immediate superordinate terms found in the definitions and the arrows stand for the classical subsumption relation [106] – demonstrates that although the individual definitions follow the Aristotelian form '*an A is a B which C*', the defined terms do not lead all together to a complete directed graph with an overarching top, not even if all 29 IASP terms would be included.

Acronym	Term (plus meaning)	N
CONC	Normal case	9
ALLO-D	**allodynia (definition)**: unexpected evoked pain	10
ALLO-N	**allodynia (note)**: unexpected more intense evoked pain	30
ANAL	**analgesia**: unexpected absence of evoked pain	40
DYS-E	**evoked dysesthesia**	80
DYS-EP	painful evoked dysesthesia	50
DYS-EU	non-painful evoked dysesthesia	30
HYPERA	**hyperalgesia**: unexpected more intense evoked pain	20
HYPERE	**hyperesthesia** = increased sensitivity to stimulation	81
HYPERE-I	unexpected more intense evoked sensation	42
HYPERE-IP	unexpected more intense evoked pain	20
HYPERE-IM	unexpected more intense evoked modus M	26
HYPERE-P	unexpected presence of evoked sensation	49
HYPERE-PU	unexpected evoked unpleasant sensation other than pain	30
HYPERE-PP	unexpected pain	10
HYPERE-PM	unexpected modus M	13
HYPERP	**hyperpathia**	30
HYPOALG	**hypoalgesia**	20
HYPOE	**hypoesthesia** = decreased sensitivity to stimulation	58
HYPOE-P	decreased sensitivity to pain stimulation	40
HYPOE-PL	less pain to pain stimulation	20
HYPOE-PA	non painful unpleasant response to pain stimulation	20
HYPOE-M	decreased sensitivity to modus M stimulation	26
HYPOE-BI	decreased sensitivity to both kinds of stimulation	8
PAR-D-E	**evoked paresthesia (definition)**	81
PAR-D-EP	painful evoked paresthesia	30
PAR-D-EU	non-painful unpleasant evoked paresthesia	30
PAR-D-EN	non-painful not unpleasant evoked paresthesia	39
PAR-N-E	**evoked paresthesia (note)**	19

Table 5 - Terms and ontological subgroups for the IASP pain assessment terminology. Legend: N = number of stimulus/response combinations applicable (max = 130).

Figure 1 - **IASP pain assessment terminology hierarchy**



Furthermore, the terms '*allodynia'* and '*hyperalgesia'* have superordinate terms which under their standard meanings should represent disjoined classes: although sensation and sensitivity are certainly related, nothing which **is** a kind of one can also be a kind of the other. In addition, already a superficial reading of these terms and accompanying notes reveals ambiguities and inconsistencies. The definition of '*allodynia'*, for instance, indicates that the term should be used for pain evoked after applying a stimulus which is below the normal pain threshold. The corresponding note however suggests that also a response on an above-threshold stimulus may count as such when the stimulus leads to more pain than expected. The note for '*dysesthesia'*, as many similar notes for other terms which for space reasons are not reproduced in Table 1 but can be found in reference [49], indicate that there is considerable overlap between the terms.

	CONC	/CONC	ALLO-D	/ALLO-D	ALLO-N	/ALLO-N	ANAL	/ANAL	DYS-E	/DYS-E	HYPERA	/HYPERA	HYPERE	/HYPERE	HYPERP	/HYPERP	HYPOALG	/HYPOALG	HYPOE	/HYPOE	PAR-D-E	/PAR-D-E	PAR-N-E	/PAR-N-E
CONC	9	0																						
/CONC	0	121																						
ALLO-D	0	10	10	0																				
/ALLO-D	9	111	0	120																				
ALLO-N	0	30	10	20	30	0																		
/ALLO-N	9	91	0	100	0	100																		
ANAL	0	40	0	40	0	40	40	0																
/ANAL	9	81	10	80	30	60	0	90																
DYS-E	0	80	10	70	30	50	20	60	80	0														
/DYS-E	9	41	0	50	0	50	20	30	0	50														
HYPERA	0	20	0	20	20	0	0	20	20	0	20	0												
/HYPERA	9	101	10	100	10	100	40	70	60	50	0	110												
HYPERE	0	81	10	71	30	51	26	55	66	15	20	61	81	0										
/HYPERE	9	40	0	49	0	49	14	35	14	35	0	49	0	49										
HYPERP	0	30	10	20	30	0	0	30	30	0	20	10	30	0	30	0								
/HYPERP	9	91	0	100	0	100	40	60	50	50	0	100	51	49	0	100								
HYPOALG	0	20	0	20	0	20	0	20	20	0	0	20	6	14	0	20	20	0						
/HYPOALG	9	101	10	100	30	80	40	70	60	50	20	90	75	35	30	80	0	110						
HYPOE	0	58	2	56	6	52	24	34	48	10	4	54	34	24	6	52	20	38	58	0				
/HYPOE	9	63	8	64	24	48	16	56	32	40	16	56	47	25	24	48	0	72	0	72				
PAR-D-E	0	81	10	71	30	51	26	55	66	15	20	61	81	0	30	51	6	75	34	47	81	0		
/PAR-D-E	9	40	0	49	0	49	14	35	14	35	0	49	0	49	0	49	14	35	24	25	0	49		
PAR-N-E	0	19	0	19	0	19	14	5	0	19	0	19	9	10	0	19	0	19	6	13	9	10	19	0
/PAR-N-E	9	102	10	101	30	81	26	85	80	31	20	91	72	39	30	81	20	91	52	59	72	39	0	111

Table 6 - Positive/negative contingency table for traditional pain terminology. A color coding is used for the 2-by-2 contingency tables to highlight the type of overlap: white indicates a symmetric overlap for all 4 types of co-occurrence; green indicates mutual exclusion of the positive occurrences, the other three colors indicate an asymmetric overlap.

12.2.2 Traditional pain assessment terminology shows considerable overlap

All terms of Table 1 could be mapped to the stimulus/response combinations. Table 6 illustrates how the parent terms relate to each other in function of the stimulus/response combinations. The individual cells contain the counts for the overlap, if any. For example, the overlap cells between hyperesthesia and hypoalgesia show - surprisingly - that these two conditions do not exclude each other: 6 of the 130 combinations fall under both definitions, 14 are such that hypoalgesia is present without hyperesthesia, 75 have hyperesthesia without hypoalgesia, and 35 don't exhibit either. An additional color coding is used to highlight the type of overlap: white indicates a symmetric overlap for all 4 types of co-occurrence as exemplified by the

hyperesthesia/ hypoalgesia pair; green indicates mutual exclusion of the positive occurrences, the other three colors indicate an asymmetric overlap. An ideal terminology would be such that the classes defined are mutually disjoint. For 12 (n) classes as is the case here, there are 66 possible overlaps (n*(n-1)/2) between any pair of these classes, not counting overlap of a class with itself. As displayed in Table 6, there is no overlap in only 2 cases of these 66: (1) for hyperpathia versus allodynia (taking the note into account), and (2) for hyperesthesia and paresthesia (when the note is not taken into account).

12.3 Novel terminology with less overlap

Table 7 provides an overview of the proposed terminology which uses 6 variables (*Response expectation*, *Main finding*, *Sensation expectation*, *Sensation intensity*, *Sensation mode*, and *Stimulation type*) that can take a number of values and which are strongly related to the variables and values used to design the analysis framework of the 130 stimulus/response combinations.

	Response expectation	**Main finding**	**Sensation expectation**	**Sensation intensity**	**Sensation mode**	**Stimulation type**		
	Concordant Discordant	***Absence Presence Configuration***	***Concordant Discordant***	***hypOresponsive hypErresponsive***	***Modal Unpleasant Painful***	***Sensation***	***Subthreshold Pain-specific Modus-specific Bimodal***	***Stimulation***
CA---SS	C	A				S	S	S
DPDEMSS	D	P	D	E	M	S	S	S
DPDEUSS	D	P	D	E	U	S	S	S
DPDEPSS	D	P	D	E	P	S	S	S
DC---SS	D	C				S	S	S
CA—MSP	C	A			M	S	P	S
CPC-PSP	C	P	C		P	S	P	S
CC---SP	C	C				S	P	S
DA—PSP	D	A			P	S	P	S
DPDOUSP	D	P	D	O	U	S	P	S
DPDOPSP	D	P	D	O	P	S	P	S
DPDEMSP	D	P	D	E	M	S	P	S
DPDEPSP	D	P	D	E	P	S	P	S
DC---SP	D	C				S	P	S
CA—USM	C	A			U	S	M	S
CPC-MSM	C	P	C		M	S	M	S
CC---SM	C	C				S	M	S
DA—MSM	D	A			M	S	M	S
DPDEMSM	D	P	D	E	M	S	M	S
DPDEUSM	D	P	D	E	U	S	M	S
DPDEPSM	D	P	D	E	P	S	M	S
DC---SM	D	C				S	M	S
CPC-MSB	C	P	C		M	S	B	S
CPC-PSB	C	P	C		P	S	B	S
CC---SB	C	C				S	B	S
DA—MSB	D	A			M	S	B	S
DA—PSB	D	A			P	S	B	S
DPDOMSB	D	P	D	O	M	S	B	S
DPDOUSB	D	P	D	O	U	S	B	S
DPDOPSB	D	P	D	O	P	S	B	S
DPDEMSB	D	P	D	E	M	S	B	S
DPDEPSB	D	P	D	E	P	S	B	S
DC---SB	D	C				S	B	S

Table 7 - Proposed alternative terminology

The values for *sensation mode* are to be interpreted as follows: 'modal' means that there is only a modal response which is not unpleasant or painful, 'unpleasant' means that the response is unpleasant but not painful, irrespective of whether there is a modal response as well, whereas 'painful' means there is only a painful response. 'Subthreshold' for *stimulation type* reflects a subthreshold stimulation for both pain and modus M, while 'bimodal' indicates an above threshold stimulation for both modus M and pain.

As is the case for the analysis framework, some values are constrained by the values for some other variables. As an example, when the value for *stimulus intensity* is 'subthreshold', there is either (1) no response in which case the value for *response expectation* is constrained to 'concordant', the value for *main finding* to 'absence', and all other variables have no value, or (2) a response is present, in which case the values for *response expectation* and *sensation expectation* are both constrained to 'discordant', the value for *main finding* to 'presence', and the value for *sensation intensity* to 'hyper-responsive'. The constraints make once again the total number of possibilities lower than can be expected: 26, excluding the combinations with the value 'configuration' for main finding which are constructed by the boolean AND-ing and OR-ing of concordant and discordant situations. The terms for this terminology are then all of the form '(*Response expectation*) (*Main finding*) ***of*** (*Sensation expectation*) (*Sensation intensity*) (*Sensation mode*) ***sensation after*** (*Stimulation type*) ***stimulation***' whereby the variables in italics are replaced by the terms for the allowed values, and the words in bold are constant. As an example, the terms for the first two combinations in Table 7 are respectively 'concordant absence of sensation after subthreshold stimulation' and 'discordant presence of discordant hyper-responsive modal sensation after subthreshold stimulation'.

The left column of Table 7 contains for further reference in Table 8 acronyms for the various possibilities formed by means of the concatenation of the individual values for a certain variable, excluding, for space reasons, the last (constant) 'S' for 'Stimulation'.

Table 8 shows the extent to which the proposed terminology categories suffer from a far less degree of overlap, overlap being indicated by the cells in light and dark red background: only 23 overlaps of the total possible 325.

12.4 Discussion

Our results in Table 5, combined with Table 1, clearly indicate that the traditional terminology is based on rather ambiguous definitions and application recommendations some of which lead to interpretations for which it is not clear whether they are intended or not. This is overwhelmingly obvious for the terms '*hyperesthesia'*, '*hypoesthesia'* and '*paresthesia'*. The latter is very broadly defined as an abnormal sensation, without making it explicit what '*abnormal'* exactly means: '*abnormal'* may indeed be interpreted as anything what is not expected, such as more or less intense pain than expected after giving a supra-threshold pain stimulus, or more or less intense pressure sensation than expected when giving a supra-threshold pressure stimulus.

It may also be interpreted as feeling an itch - a form of unpleasant sensation - when giving a pressure stimulus with or without there being a pressure sensation, and so forth. The note for paresthesia, in contrast, tells us that only '*not unpleasant'* sensations should count as qualifying, which limits the number of possibilities considerably.

It leaves however still many interpretations open, such as whether the resulting sensation must be alien to the given stimulus - would an erotic feeling induced by providing a pressure stimulus to the hand count as such a non-unpleasant abnormal sensation? - or whether it may be special cases of hypo- and hyperesthesia.

These reflections provide at the same time explanations for the very high degree of overlap between the majority of the traditional terms (Table 6). There is of course a symmetric non-overlap for each category with each negation, but the only non-overlap between distinct categories is found for the pairs allodynia (taking the note into account) -hyperpathia and hyperesthesia-paresthesia (as defined, without the limiting note).

The proposed terminology shows a much more limited degree of overlap. This lesser degree of overlap is because the parameters have been chosen in such a way that a specific combination of values cannot count for a specific class in more than one way, a feature which is not exhibited by the traditional terminology.

A disadvantage of the terminology is that it is more verbose, but this is compensated by the ease by which it can be implemented in systems for structured electronic reporting and automatic assigning of the categories using single select choice lists for each variable.

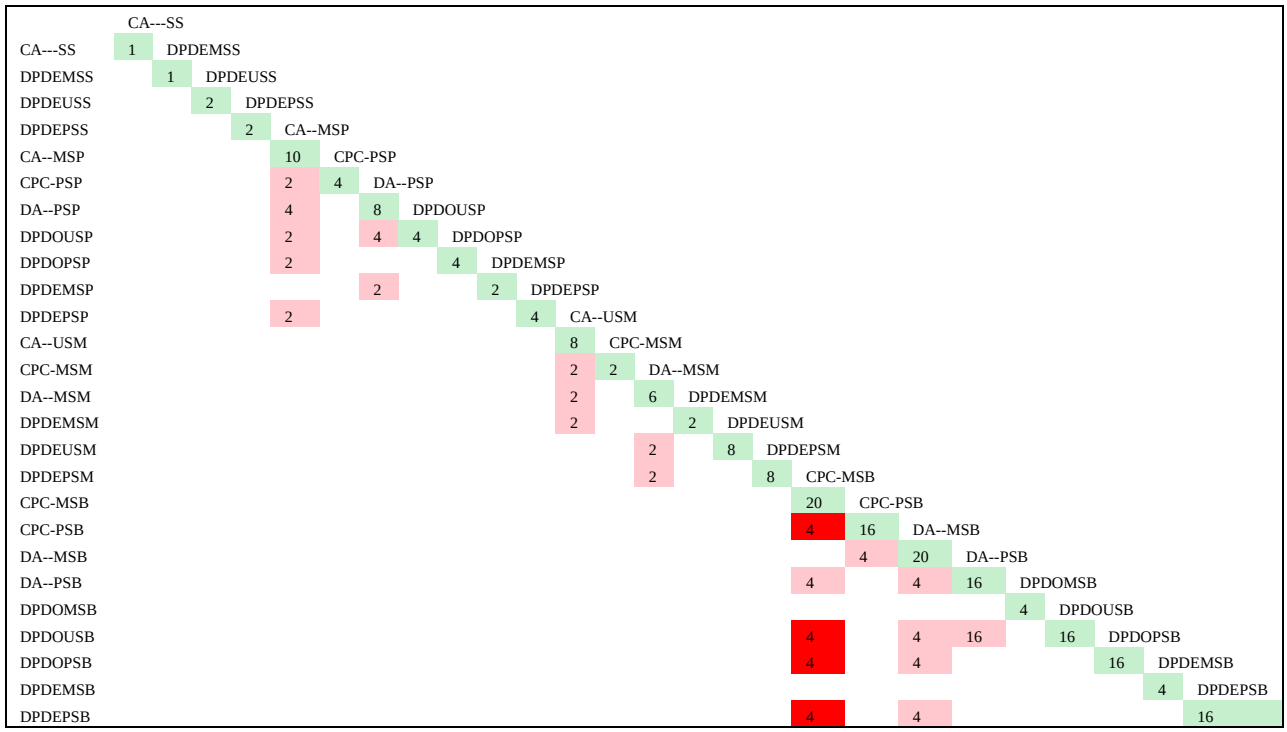


Table 8 - Overlap between proposed pain assessment categories

12.5 Conclusion

It is demonstrated that the IASP terms do not satisfy the criteria for direct integration in a realism-based ontology. A new terminology for stimulus based pain and somatosensory status assessment is proposed which exhibits less shortcomings in terms of overlap than the traditional terminology. This is because in contrast to the traditional approach, this proposal does not underestimate the various stimulus/response combinations that may occur.

13 Ontological perspectives on biomarkers and diagnostic classifications for orofacial pain disorders

The Institute of Medicine defines a biomarker as, '... characteristics that are objectively measured and evaluated as indicators of normal biological processes, pathogenic processes or responses to an intervention.' For the notion of biomarker to play a prominent role in diagnostic classifications, for instance in the formulation of diagnostic criteria, there must be a uniform understanding amongst developers of such classifications about what biomarkers precisely are and whether all entities to which the term 'biomarker' is assigned form a uniform group. For a group of entities to be uniform, all and only its members must exhibit a certain combination of characteristics and this is so irrespective of whether science has advanced enough to discover this unique combination of characteristics and whether an appropriate terminology has been developed to report on these characteristics adequately. This understanding must thus also include in what way biomarkers are distinct from other entities on the side of the patient such as signs, symptoms, diagnostic tests that are applied to them, all entities which are already standardly referred to in the formulation of diagnostic classes and corresponding criteria. And finally, a similar understanding must be established for each of the various sorts of biomarkers as, for instance, suggested by terms used in previous sections such as 'investigative biomarkers', 'prognostic biomarkers', 'radiologic biomarkers', and so forth.

Ontological theories are instrumental in reaching such understanding and in documenting the insights gained [79]. Unfortunately, not much careful attention has thus far been paid to the ontological status of what biomarkers are: of the 370 biomedicine oriented representational artifacts accessible through the BioPortal of the National Center for Biomedical Ontology (NCBO) [91], only 10 have a representational unit for 'biomarker' (Figure 1) and the taxonomies in which this class appears vary widely. Is it the vagueness of the term 'characteristic' – this term does not denote an acceptable ontological category at all – for that what according to the IOM a biomarker is supposed to be which let the authors of these 10 representational artifacts go in different directions in their taxonomy development? Is it because these authors did not follow coherent ontological principles? Or has the notion of 'biomarker' not yet drawn enough attention from skilled biomedical ontologists to enjoy a widely accepted interpretation? It is most likely a combination of all three. Six out of the ten artifacts exhibit serious deficiencies for the following reasons: (1) the suggested top category does not correspond to the most generic entity in reality: '1', '2', '3', '5', and '10'; (2) the erroneous belief that biomarkers, and more general, organism attributes are pure conceptual: '9'. The taxonomies of the four other representational artifacts – '4', '6', '7' and '8' – suggest that their authors attempted to provide an interpretation of what the term 'biomarker' might denote in terms of the principles of ontological realism. This may be inferred from the biomarker class in these representational artifacts being subsumed by 'specifically dependent continuant' which is a class coming from the Basic Formal Ontology [107] which implements ontological realism. Representational artifact '7' includes this class via the class 'quality' as immediate subclass and the three others via 'role'. As will be demonstrated, these views exhibit shortcomings as well, but, nevertheless, the analysis of them provides useful insight into what the research agenda for pain specialists with respect to biomarkers and their inclusion in diagnostic classifications ought to be.

13.1 Biomarkers as roles

The authors of the representational artifacts labeled '2', '3', '4', '5', and '6' entertain a view according to which not the entities that in medical discourse are denoted by the term 'biomarker' are instances of the ontological category biomarker, but rather the *role* played by some of these

entities. That role would then be to serve as indicators for one or other phenomenon in line with the IOM's definition. A similar view is entertained for well-known roles in healthcare such as those of patient and clinician. It is indeed the case that the words 'patient' and 'clinician' are in medical discourse much more often used to designate a person in which inheres the patient, resp. clinician role, rather than these roles themselves. Similarly, the word 'biomarker' is in medical discourse primarily used to denote entities in which inheres the biomarker role, rather than that role itself. Interestingly, none of the representational artifacts of which the relevant parts are depicted in Figure 1 use 'biomarker' in the medical discourse sense.

The Basic Formal Ontology describes *roles* as *realizable entities* that exist because there is some bearer that is in some special physical, social, or institutional set of circumstances in which this bearer does not have to be and which is not such that, if it ceases to exist, then the physical make-up of the bearer would thereby be changed. An entity has thus a certain role not because of the way it itself is, but because of something that happens or obtains externally. Clearly, when certain human beings start or stop to be patient or clinician there need not be any change in the physical make-up of these human beings. These roles are assigned to them: the role of clinician to human beings that satisfy the societal requirements for providing certain types of care, and the role of patient to human beings that enter into a care receiving relationship with clinicians. Granted, human beings in the role of patient do have usually a physical make-up – some disorder – which is different from healthy human beings, but it is not in virtue of that distinct make-up that they are patients: they become patients when they seek care, and that is even so when they have no disorder at all. Similarly, there need not be any changes in the physical make-up of bodily components to become qualified – or disqualified – as, for instance, molecular biomarkers. If there were changes that led molecules such as MCP-1 and IL-8 to become biomarkers, then that were changes in our ability to measure these components more reliably and in our understanding on how they relate to certain types of disorder, changes in reality which are all changes external to the bodily components.

The view that biomarkers – in the medical discourse sense – can play a biomarker role – in the ontological sense – can be argued to line nicely up with the differentiae for biomarkers posited by the IOM: roles being realizable entities means that the entities in which inhere a role can participate in processes which realize that role. When a clinician investigates a patient, that process of investigating that patient is a realization of the clinician role. When a pain researcher examines the MCP-1 concentration in a TMD patient because MCP-1 is recognized as biomarker for local inflammation, then that very measurement is the realization of the biomarker role assigned to the collection of MCP-1 molecules in that patient.

There are nevertheless certain shortcomings with this view, one being that it cannot account for all phenomena the IOM had most likely in mind. Indeed, contending that the word 'biomarker' denotes a role in the ontological realist sense comes down to arguing that the bearer of that role is an *independent continuant*, i.e. an entity that does not need – in technical terms does not 'depend on' – another entity for its ontological existence. Examples in the domain of health are not only human beings and their body parts ranging from bodily systems and organs to cells, cellular components and molecules and molecular complexes (including neurotransmitters and nociceptors), but also medical devices and concretized data such as notes in paper-based or electronic health records or spreadsheets. Amongst the biomedical entities that are not independent continuants are for instance *processes* (pain sensation, inflammatory processes, neurotransmission, central pain modulation, …), and dependent continuants such as *qualities* (blood pressure, skin color, …) and *functions* (e.g. the function of certain nerves to transmit sensory signals). These entities can thus not enjoy a biomarker role! A solution here would be to carry out further investigations into the types of independent continuants upon which these

dependent entities depend and then to assess whether the biomarker role can be assigned to them. This is not an enterprise to be conducted by ontologists but rather by domain experts; in the case of pain research this would be neurophysiologists, neuroscientists, and so forth.

13.2 Biomarkers as qualities

Qualities, according to the Basic Formal Ontology, are specifically dependent continuants that do not require any further process in order to be realized: the shape of, for example, the temporal mandibular joint of some specific patient exists without the need for any process in which that TMJ might participate. Specific clinically abnormal TMJ shapes may bring into being the disposition to produce clicking noises when the mouth is opened. Whereas the disposition to produce clicking noises requires opening the mouth to be realized, the existence of the shape of the TMJ does not require that process, or any other process, to exist, and this is so despite the fact that the configuration of the shape changes with each different position of the jaw: changes in the shape of the TMJ do not bring that shape in or out of existence.

Entertaining a view according to which being a biomarker is a quality rather than a role means putting the emphasis on the status of a biomarker as an indicator for some phenomena instead of on the measurability and its status of being selected for a given reason such as the objective measurability. Whereas measurability of, for example, MCP-1 concentration is a realizable entity, and measurements of MCP-1 concentrations are realizations of that measurability which happen only when performed, the MCP-1 concentration in any given patient is what it is at all times, whether or not it is measured, and whether or not the actual concentration changes from time to time. Also, that MCP-1 increases in case of local inflammation, and that therefore MCP-1 concentration is an indicator for local inflammation, was already the case before human beings had any idea about inflammation, before MCP-1 was discovered, and before the relationship between MCP-1 and local inflammation was discovered. Mere lack of knowledge does not change what is the case in reality. This is an argument in favor of the view that at least some types of entities are associated with a biomarker quality rather than a biomarker role. We deliberately wrote 'some types' because as with roles, qualities cannot depend on processes and thus cannot provide the complete picture

13.3 Biomarkers and the Ontology of General Medical Science

The Ontology of General Medical Science (OGMS) is based on a terminological framework that encompasses diseases, their causes and manifestations, and diagnostic acts and other entities pertaining to the ways diseases are recognized and interpreted in the clinic. The framework was designed to avoid the sort of conflations often encountered in medical discourse between entities on the side of the organism – in the case of healthcare: human beings – and the *evidence* for the existence of such entities [31]. This and other conflations are widespread, and it is thus no surprise to find examples thereof in the IOM's report on biomarkers, for instance in '*Cholesterol and blood sugar levels are biomarkers, as are blood pressure, enzyme levels, measurements of tumor size from MRI or CT, and the biochemical and genetic variations observed in age-related macular degeneration.*' [1, p2] where characteristics on the side of the patient are conflated with measurements of these characteristics, it even being unclear whether by 'measurements' is meant either (a) the processes of measuring an entity on the side of the patient or (b) the data – usually expressed as values of some sort – obtained through such a process of measuring. In [1, p18], the need is expressed '*…to develop a transparent process for creating well-defined consensus standards and guidelines for* ***biomarker development, validation, qualification, and use*** [bold emphasis added] *to reduce the uncertainty in the process of development and adoption*'. This would restrict biomarkers to be measuring processes and/or devices to assist in such processes as it is hard to fathom that what is

proposed to be developed here are blood sugar levels and tumor sizes. But that then, in turn, cannot be lined up with the IOM's definition for biomarker which is stated to be something that is (a) objectively measured – surely, the idea is not that what is measured would be the measuring process of, for instance, blood glucose itself – and (b) an indicator for normal, pathological or response to treatment processes – clearly, the mere performance of some test is itself not an indication at all of what is going on in the patient, rather an indication of what is going on in the mind of the clinician as he is trying to find out what is going on in the patient.

That the terminology around biomarkers is inconsistent – a problem the IOM recognizes in its own report [1, p22] but unfortunately is contributing to rather than solving it – does not mean that the ideas behind it don't have value. But it does mean that the terminology needs to be rendered unambiguous and anchored in an ontology which recognizes all types of entities to be referred to in standards and guidelines for biomarker development, validation, qualification, and use. For this, the OGMS is an ideal candidate.

The basic axiom of the OGMS is that every disease rests always on some (perhaps as yet unknown) physical basis. When, for example, there is in a specific patient an elevated level of TNF in the synovial fluid of the TMJ, then this is because (1) some physical structure or substance in the organism is disordered (e.g. physical damage of some sort in the TMJ – the *disorder*) as a result of which (2) there exists a disposition (e.g. TMD of some sort – the *disease*) for the organism to act in a certain abnormal way. This disposition – another type of realizable entity as described above – in question is realized by pathological processes (e.g. inflammation) including manifestations that can be recognized as symptoms and signs of the disorder (e.g. pain, clicking noises, decreased mobility) or through measurement assays (e.g. laboratory tests, imaging procedures). The core definitions for entities on the side of the patient arising from this view are (the terms in bold are to be interpreted in the strict technical sense as defined, and the definitions are not an attempt to describe how these terms are used in medical discourse at every single occasion):

- **Etiological Process =def. –** *A process in an organism that leads to a subsequent disorder.*
- **Disorder =def. –** *A causally relatively isolated combination of physical components that is (a) clinically abnormal and (b) maximal, in the sense that it is not a part of some larger such combination.*
- **Pathological Process =def.** *– A bodily process that is a manifestation of a disorder*.
- **Disease =def. –** *A disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.*
- **Disease Course =def. –** *The totality of all processes through which a given disease instance is realized.*
- **Manifestation of a Disease =def. –** *A bodily feature of a patient that is (a) a deviation from clinical normality that exists in virtue of the realization of a disease and (b) is observable. Observability includes observable through elicitation of response or through the use of special instruments.*

- **Phenotype =def. –** *A (combination of) bodily feature(s) of an organism determined by the interaction of its genetic make-up and environment.*
- **Clinical Phenotype =def. –** *A clinically abnormal phenotype.*
- **Disease Phenotype =def. –** *A clinical phenotype that is characteristic of a single disease.*

Entities that qualify as bodily features are (1) physical components such as bodily components (nerve cells, nociceptors, neurotransmitters, etc.) and external components (pathogens, toxins, microbiome, …), (2) bodily qualities such as cytokine concentrations and (3) bodily processes in which physical components participate, irrespective of them being normal (neurotransmission and concordant pain sensation), pathological (phantom pain), or induced through interventions.

A **disease phenotype** can exist without being observed. With the advance of technology, the ability to detect more underlying components will expand. The **clinical phenotype** – for a specific patient – incorporates the abnormal **phenotypes** realized at each stage of the **disease course**. A **disease phenotype** may be a single type of abnormality characteristic of a given disease; or a combination of several **manifestations of a disease** and clinically normal physical components, ordered in a temporal sequence characteristic of one or more typical **disease courses** for the given **disease**.

Core definitions for entities that are the result for what is observed, are:

- **Clinical Picture =def. –** *A representation of a clinical phenotype that is inferred from the combination of laboratory, image and clinical findings about a given patient.*
- **Diagnosis =def.** *– A conclusion of an interpretive process that has as input a clinical picture of a given patient and as output an assertion to the effect that the patient has a disease of such and such a type.*

The view on biomedical reality offered by the OGMS allows us to replace the vague term ‘characteristics’ in the IOM’s definition for biomarker by ‘bodily features’, at least under the assumption that the IOM intended biomarkers to be entities on the side of the patient, and not investigative processes to determine the nature of these entities, nor the data obtained through these processes.

> **Biomarker =def. –** *A bodily feature which is objectively measurable and of which the existence is the result of some normal biological process, of some pathogenic process or of some response to an intervention.*

With this as starting point, further types of biomarkers can be defined in very precise terms, e.g.:

> **Disease Biomarker =def. –** *A **manifestation of a disease** which is objectively measurable and which is part of a **disease phenotype**.*
>
> **Diagnostic Biomarker =def. –** *A **disease biomarker** of which a representation is part of a **clinical picture** which serves as input for a **diagnosis**.*

13.4 Recommendations for ontology-based representation of biomarkers in diagnostic classifications and related criteria for orofacial pain

Both research aiming the discovery of suitable biomarkers for orofacial pain and the adequate use thereof to build diagnostic classifications will benefit from the advantages ontology has to offer. A good start will be to clean up the terminology around biomarkers the scope of which is must broader than the domain of pain research. If this is not immediately feasible for biomedicine in general, then at least pain researchers could get a competitive advantage by implementing a few simple steps.

A good start would be to develop on the basis of the literature an inventory of biomarker candidates relevant for pain research and subsequent application for diagnostics. This inventory should include for each biomarker a number of essential information elements. One is the type of bodily feature the biomarker is an instance of which needs to be at least expressed in terms of the OGMS, the most generic allowed types being physical component, bodily process and bodily quality. If the biomarker is determined to be a physical component, then further subtyping should be documented using ontologies accepted in the Open Biomedical Ontologies Foundry or candidate ontologies thereof, examples being the Foundational Model of Anatomy for any bodily component down to individual cells, the Cellular Component taxonomy of the Gene Ontology, the Protein Ontology, and so forth.

If the biomarker is a bodily process, good candidates are the Biological Process and Molecular Function taxonomies of the Gene Ontology, with respect to the latter on the condition that the biomarker is documented as being a process which realizes a given molecular function. Since bodily processes always depend on at least one bodily component, it should for such a biomarker also be indicated which bodily components it depends on, using one of the ontologies just mentioned. If that is not documented in the ontology used to type the biomarker, it should be added at the level of the inventory.

If the biomarker is a bodily quality, a good ontology for further subtyping is the Phenotypic Quality Ontology (PATO). As with processes, the inventory should further contain information about what bodily component this biomarker is a quality of.

The second sort of information the inventory should contain, is the type of investigation that in the cited literature source is used to determine the biomarker being documented. A candidate ontology for this is the Ontology of Biomedical Investigations (OBI). Since varies types of assays can be used to measure the same biomarker, there might need to be several distinct measurement related entries for each biomarker.

The third piece of information is, in case of an inventory aimed towards diagnostic classification development, the underlying pathology for which the biomarker is believed to be an indicator. This information is typically available as a diagnosis. To be fully in line with the OGMS, this diagnosis should be at least brought in closer relation with a disorder, a disease and/or a disease course as described above.

14 OPMQoL Upper Ontology

The goal of the OPMQoL is to make it possible to describe datasets obtained from pain research in a uniform and formal way, and that is general enough to include other datasets in the same domain once they become available. The importance of this endeavor lays in its contribution to solving an important problem, namely that the phenotype of all orofacial pain conditions is insufficiently defined in terms of the scope, the natural history and/or clinical course of the disease subgroup of interest, and, most importantly, with respect to disease traits for which laboratory research has provided important pathogenetic insight [38].

The ontology is being build and continuously updated following the principles adhered to in the Open Biomedical Ontology Foundry (OBO-Foundry) [39], using Basic Formal Ontology (BFO) [40], and Referent Tracking (RT) [41] as generic semantic technologies.

14.1 Feeder Ontologies

14.1.1 The Basic Formal Ontology (BFO/BFO2)

Numerous domain ontologies use the *Basic Formal Ontology* (BFO) as an upper level reference ontology. BFO is a realist, formal and domain-neutral upper level ontology that is designed to represent at a very high level of generality the types of entities that exist in the world and the relations that hold between them [108-110]. BFO is intended to provide the most basic building blocks for the construction of domain-specific ontologies at lower levels. Briefly, it provides a starting point for logical descriptions (formulated through the statement of necessary and jointly sufficient conditions) of the types of entities in a specific domain. Because of this common starting point, the domain ontologies using BFO appropriately are to a degree interoperable.

14.1.2 The Ontology of General Medical Science (OGMS)

The Ontology of General Medical Science (OGMS) provides a collection of carefully defined representational units that allow biomedical researchers to describe and classify what they observe in terms of, for instance, *disorders*, *diseases*, *diagnoses*, *clinical pictures*, and so forth, or, not less important, identify where terminology as currently used goes astray [31]. This methodology allowed, for example, to distinguish six types of pain-related phenomena implicitly present in the IASP definition for ‘pain’ [97] and to provide an ontologically adequate description of what is called ‘persistent dento-alveolar pain disorder’ (PDAP) [3].

14.1.3 The Mental Functioning/Emotion Ontology (EMO/OMD)

The Mental Functioning Ontology covers entities such as perceptions, beliefs, emotions and desires, which in line with OGMS have a physical basis (in the brain and perceptual organs), in the relevant components of which there occur processes of certain sorts such as: activations of neurons, formation of synapses between cells, flows of electrons, and so forth. The corresponding physical components in the patient organism – components which are involved in both mental disease and normal cognitive functioning – are called ‘mental functioning related anatomical structures’ [4].

14.1.4 The Information Artifact Ontology (IAO)

This ontology under development represents various sorts of information types (data items, report requests, measurement results, etc.) as subtypes under BFO’s generically dependent continuant type.

14.2 Latest development version

The following table gives of an overview of the top-domain types under which the entities from the processed datasets and assessment instruments are classified.

Nr	Type	Source	Definition
1	entity	BFO	
2	-continuant	BFO	An entity [bfo:Entity] that exists in full at any time in which it exists at all, persists through time while maintaining its identity and has no temporal parts.
3	--dependent continuant	BFO	A continuant [snap:Continuant] that is either dependent on one or other independent continuant [snap:IndependentContinuant] bearers or inheres in or is borne by other entities.
4	---generically dependent continuant	BFO	A continuant [snap:Continuant] that is dependent on one or other independent continuant [snap:IndependentContinuant] bearers. For every instance of A requires some instance of (an independent continuant [snap:IndependentContinuant] type) B but which instance of B serves can change from time to time.
5	----ICE	IAO	an information content entity is an entity that is generically dependent on some artifact and stands in relation of aboutness to some entity
6	-----CRID	IAO	An information content entity that consists of a CRID symbol and additional information about which CRID registry it belongs.
7	-----data item	IAO	a data item is an information content entity 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.
8	------dataset-record	IAO-proposal	
9	------cartesian spatial coordinate datum	IAO	A cartesian spatial coordinate datum is a representation of a point in a spatial region, in which equal changes in the magnitude of a coordinate value denote length qualities with the same magnitude
10	-------one dimensional cartesian spatial coordinate datum	IAO	NA
11	-------two dimensional cartesian spatial coordinate datum	IAO	NA
12	-------three dimensional cartesian spatial coordinate datum	IAO	NA

13	------clinical finding	OGMS	A representation that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof.
14	-------clinical history	OGMS	A series of statements representing health-relevant qualities of a patient and of a patient's family.
15	-------image finding	OGMS	A representation of an image that supports an inference to an assertion about some quality of a patient.
16	-------laboratory finding	OGMS	A representation of a quality of a specimen that is the output of a laboratory test and that can support an inference to an assertion about some quality of the patient.
17	-------physical examination finding	OGMS	NA
18	------clinical picture	OGMS	A representation of the clinically significant bodily components and/or bodily processes of a human being that is inferred from the totality of relevant clinical findings.
19	------data about an ontology part	IAO	data about an ontology part is a data item about a part of an ontology, for example a term
20	-------DbXref	IAO	NA
21	-------Definition	IAO	NA
22	-------Subset	IAO	NA
23	-------Synonym	IAO	NA
24	-------SynonymType	IAO	NA
25	-------curation status specification	IAO	The curation status of the term. The allowed values come from an enumerated list of predefined terms. See the specification of these instances for more detailed definitions of each enumerated value.
26	-------denotator type	IAO	A denotator type indicates how a term should be interpreted from an ontological perspective.
27	-------obsolescence reason specification	IAO	The reason for which a term has been deprecated. The allowed values come from an enumerated list of predefined terms. See the specification of these instances for more detailed definitions of each enumerated value.
28	------data set	IAO	A data item that is an aggregate of other data items of the same type that have something in common. Averages and distributions can be determined for data sets.
29	-------CRID Registry	IAO	A CRID registry is a dataset of CRID records, each consisting of a CRID symbol and additional information which was recorded in the dataset through a assigning a centrally registered identifier process.

30	-------time sampled measurement data set	IAO	A data set that is an aggregate of data recording some measurement at a number of time points. The time series data set is an ordered list of pairs of time measurement data and the corresponding measurement data acquired at that time.
31	------diagnosis	OGMS	The representation of a conclusion of a diagnostic process.
32	------measurement datum	IAO	A measurement datum is an information content entity that is a recording of the output of a measurement such as produced by a device.
33	-------scalar measurement datum	IAO	a scalar measurement datum is a measurement datum that is composed of two parts, numerals and a unit label.
34	--------length measurement datum	IAO	A scalar measurement datum that is the result of measurement of length quality
35	--------mass measurement datum	IAO	A scalar measurement datum that is the result of measurement of mass quality
36	--------time measurement datum	IAO	A scalar measurement datum that is the result of measuring a temporal interval
37	------preclinical finding	OGMS	A representation of a quality of a patient that is (1) recorded by a clinician because the quality is hypothesized to be of clinical significance and (2) refers to qualities obtaining in the patient prior to their becoming detectable in a clinical history taking or physical examination.
38	------prognosis	OGMS	A hypothesis about the course of a disease.
39	------setting datum	IAO	A settings datum is a datum that denotes some configuration of an instrument.
40	-----directive information entity	IAO	An information content entity whose concretizations indicate to their bearer how to realize them in a process.
41	------action specification	IAO	a directive information entity that describes an action the bearer will take
42	------conditional specification	IAO	a directive information entity that specifies what should happen if the trigger condition is fulfilled
43	-------rule	IAO	a rule is an executable which guides, defines, restricts actions
44	-------time trigger	IAO	NA
45	------data format specification	IAO	A data format specification is the information content borne by the document published defining the specification. Example: The ISO document specifying what encompasses an XML document; The instructions in a XSD file
46	------objective specification	IAO	a directive information entity that describes an intended process endpoint. When part of a plan specification the concretization is realized in a planned process in which the bearer tries to effect the world so that the process endpoint is achieved.

47	------plan specification	IAO	a directive information entity that when concretized it is realized in a process in which the bearer tries to achieve the objectives, in part by taking the actions specified. Plan specifications includes parts such as objective specification, action specifications and conditional specifications.
48	-------algorithm	IAO	A plan specification which describes inputs, output of mathematical functions as well as workflow of execution for achieving an predefined objective. Algorithms are realized usually by means of implementation as computer programs for execution by automata.
49	--------software interpreter	IAO	An algorithm that takes, as input, some digital entity, and takes action driven by the information content of that algorithm
50	-------programming language	IAO	A language in which source code is written, intended to executed/run by a software interpreter. Programming languages are ways to write instructions that specify what to do, and sometimes, how to do it.
51	-------software	IAO	Software is a plan specification which is a series of encoded instructions that can be directly executed by a processing unit or transformed in to a form that can be. For programming texts that are syntactically correct and which are in a language that can be executed by an interpreter this would correspond to the tokenized version of the text stripped of comments.
52	--------software application	IAO	A software application is software that can be directly executed by some processing unit.
53	--------software library	IAO	A software library is software composed of a collection of software modules and/or software methods in a form that can be statically or dynamically linked to some software application.
54	--------software method	IAO	A software method (also called subroutine, subprogram, procedure, method, function, or routine) is software designed to execute a specific task.
55	--------software module	IAO	A software module is software composed of a collection of software methods.
56	--------software script	IAO	A software script is software whose instructions can be executed using a software interpreter.
57	-------study design	IAO	A study design is a plan specification comprised of protocols (which may specify how and what kinds of data will be gathered) that are executed as part of an investigation and is realized during a study design execution.
58	------source code module	IAO	The written source code that implements part of an algorithm. Test - if you know that it was written in a specific language, then it can be source code module. We mean here, roughly, the wording of a document such as a perl script.
59	-----document	IAO	A collection of information content entities intended to be understood together as a whole
60	------patent	IAO	A document that has been accepted by a patent authority
61	------publication	IAO	A document that has been accepted by a publisher

62	-------publication about an investigation	IAO	A publication that is about an investigation
63	------report	IAO	a document assembled by an author for the purpose of providing information for the audience. A report is the output of a documenting process and has the objective to be consumed by a specific audience. Topic of the report is on something that has completed. A report is not a single figure. Examples of reports are journal article, patent application, grant progress report, case report (not patient record)
64	-------journal article	IAO	a report that is published in a journal
65	-----document part	IAO	an information content entity that is part of a document
66	------abstract	IAO	A summary of the entire document that is substantially smaller than the document it summarizes. It is about the document it summarizes.
67	------acknowledgements section	IAO	Part of a publication that is about the contributions of people or institutions other than the authors.
68	------author contributions section	IAO	A part of a publication that is about the specific contributions of each author
69	------author list	IAO	part of a document that enumerates the authors of the document
70	------copyright section	IAO	A document part that describes legal restrictions on making or distributing copies of the document
71	------discussion section of a publication about an investigation	IAO	A part of a publication about an investigation that is about the study interpretation of the investigation
72	------footnote	IAO	A part of a document that is about a specific other part of the document. Usually footnotes are spatially segregated from the rest of the document.
73	------institution list	IAO	part of a document that has parts that are institution identifications associated with the authors of the document
74	------introduction to a publication about an investigation	IAO	A part of a publication about an investigation that is about the objective specification (why the investigation is being done)
75	------methods section	IAO	A part of a publication about an investigation that is about the study design of the investigation
76	------references section	IAO	A part of a document that has citations as parts
77	------results section	IAO	A part of a publication about an investigation that is about a study design execution
78	------supplementary material to a document	IAO	part of a document that is segregated from the rest of the document due to its size
79	-----email address	IAO	NA

80	-----figure	IAO	An information content entity consisting of a two dimensional arrangement of information content entities such that the arrangement itself is about something.
81	------diagram	IAO	A figure that expresses one or more propositions
82	-------graph	IAO	A diagram that presents one or more tuples of information by mapping those tuples in to a two dimensional space in a non arbitrary way.
83	--------venn diagram	IAO	A Venn diagram is a report graph showing all hypothetically possible logical relations between a finite collection of sets.
84	--------contour plat	IAO	NA
85	--------dendrogram	IAO	A dendrogram is a report graph which is a tree diagram frequently used to illustrate the arrangement of the clusters produced by a clustering algorithm.
86	--------density plot	IAO	A density plot is a report graph which is a graphical representation of data where the tint of a particular pixel corresponds to some kind of function corresponding the the amount of data points relativelly with their distance from the the pixel.
87	--------dot plat	IAO	A dot plot is a report graph which is a graphical representation of data where each data point is represented by a single dot placed on coordinates corresponding to data point values in particular dimensions.
88	--------heatmap	IAO	A heatmap is a report graph which is a graphical representation of data where the values taken by a variable(s) are shown as colors in a two-dimensional map.
89	--------histogram	IAO	A histogram is a report graph which is a statistical description of a distribution in terms of occurrence frequencies of different event classes.
90	--------line graph	IAO	A line graph is a type of graph created by connecting a series of data points together with a line.
91	--------scatter plat	IAO	A scatterplot is a graph which uses Cartesian coordinates to display values for two variables for a set of data. The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.
92	--------survival curve	IAO	A survival curve is a report graph which is a graphical representation of data where the percentage of survival is plotted as a function of time.
93	------image	IAO	An image is an affine projection to a two dimensional surface, of measurements of some quality of an entity or entities repeated at regular intervals across a spatial range, where the measurements are represented as color and luminosity on the projected on surface.
94	-------photograph	IAO	A photograph is created by projecting an image onto a photosensitive surface such as a chemically treated plate or film, CCD receptor, etc.

95	-----narrative object	IAO	A narrative object is an information content entity that is a set of propositions.
96	-----patient symptom report	OGMS	A communication from a patient about something they perceive as being abnormal about their body or life.
97	-----symbol	IAO	a smallish, word-like datum...
98	------CRID symbol	IAO	A symbol that is part_of a CRID and that is sufficient to look up a record from the CRID's registry.
99	------lot number	IAO	A lot number is an information content entity which is an identical sequence of character borne by part of manufactured product or its packaging for each instances of a product class in a discrete batch of an item. Lot numbers are usually assigned to each separate production run of an item. Manufacturing as a lot might be due to a variety of reasons, for example, a single process during which many individuals are made from the same portion of source material. Lot numbers can be encoded in a pattern of other information objects, such as bar codes, numerals, or patterns of dots.
100	------model number	IAO	A model number is an information content entity specifically borne by catalogs, design specifications, advertising materials, inventory systems and similar that is about manufactured objects of the same class. The model number is an alternative term for the class. The manufactured objects may or may not also bear the model number. Model numbers can be encoded in a variety of other information objects, such as bar codes, numerals, or patterns of dots.
101	------numeral	IAO	A symbol that denotes a number.
102	-------integer numeral	IAO	a numeral that denotes an integer
103	------serial number	IAO	A serial number is an information content entity which is a unique sequence of characters borne by part of manufactured product or its packaging that is assigned to each individual in some class of products, and so can serve as a way to identify an individual product within the class. Serial numbers can be encoded in a variety of other information objects, such as bar codes, numerals, or patterns of dots.
104	------version number	IAO	A version number is an information content entity which is a sequence of characters borne by part of each of a class of manufactured products or its packaging and indicates its order within a set of other products having the same name.
105	-----textual entity	IAO	A textual entity is a part of a manifestation (FRBR sense), a generically dependent continuant whose concretizations are patterns of glyphs intended to be interpreted as words, formulas, etc.

106	------caption	IAO	A textual entity that describes a figure
107	------citation	IAO	a textual entity intended to identify a particular publication
108	------conclusion textual entity	IAO	A textual entity that expresses the results of reasoning about a problem, for instance as typically found towards the end of scientific papers.
109	------document title	IAO	A textual entity that names a document
110	-------running title	IAO	A shorter version of a document title
111	------hypothesis textual entity	IAO	A textual entity that expresses an assertion that is intended to be tested.
112	------institutional identification	IAO	A textual entity intended to identify a particular institution
113	------postal address	IAO	A textual entity that is used as directive to deliver something to a person, or organization
114	------table	IAO	A textual entity that contains a two-dimensional arrangement of texts repeated at regular intervals across a spatial range, such that the spatial relationships among the constituent texts expresses propositions
115	-------table of abbreviations	IAO	A table where the constituent texts are abbreviations and their expansions
116	-------table of contents	IAO	A table that relates document parts to specific locations in a document (usually page numbers). This is also a document part (subsumption there should be inferred).
117	-------table of figures	IAO	A table that relates figures in a document to specific locations in that document (usually page numbers). This is also a document part (subsumption there should be inferred).
118	------written name	IAO	A textual entity that denotes a particular in reality.
119	-------author identification	IAO	A textual entity intended to identify a particular author
120	----normal value	OGMS	A value for a quality reported in a lab report and asserted by the testing lab or the kit manufacturer to be normal based on a statistical treatment of values from a reference population.
121	---specifically dependent continuant	BFO	A continuant [snap:Continuant] that inheres in or is borne by other entities. Every instance of A requires some specific instance of B which must always be the same.
122	----cognitive representation	OMD/EMO	NA
123	-----affective representation	OMD/EMO	NA
124	------subjective emotional feeling	EMO(OMD)	NA
125	-----appraisal	OMD/EMO	[Will Hsu] the classification of someone or something with respect to its worth
126	----quality	BFO	A specifically dependent continuant [snap:SpecificallyDependentContinuant] that is exhibited if it inheres in an entity or entities at all (a categorical property).
127	-----behavioral inducing state	OMD/EMO	NA

128	-----configuration	OGMS	A quality which is an spatial arrangement or distribution of a(n) independent continuant(s) across a Three Dimensional Region.
129	------pathological configuration	OGMS	A configuration which deviates in some way from a canonical configuration for a particular organism.
130	-----information carrier	IAO	A quality of an information bearer that imparts the information content
131	-----length	IAO	A 1-D extent quality which is equal to the distance between two points.
132	-----manifestation of a disease	OGMS	A quality of a patient that is (a) a deviation from clinical normality that exists in virtue of the realization of a disease and (b) is observable.
133	------clinical manifestation of a disease	OGMS	A manifestation of a disease that is detectable in a clinical history taking or physical examination.
134	------preclinical manifestation of a disease	OGMS	A manifestation of a disease that exists prior to the time at which it would be detected in a clinical history taking or physical examination, if the patient were to present to a clinician. A realization of a disease that exists prior to its becoming detectable in a clinical history taking or physical examination.
135	-----mass	IAO	A physical quality that inheres in a bearer by virtue of the proportion of the bearer's amount of matter.
136	-----phenotype	OGMS	A (combination of) quality(ies) of an organism determined by the interaction of its genetic make-up and environment that differentiates specific instances of a species from other instances of the same species.
137	------clinical phenotype	OGMS	A clinically abnormal phenotype.
138	------disease phenotype	OGMS	A clinically abnormal phenotype that is characteristic of a single disease.
139	-----relational quality	BFO2	b is a relational quality = Def. for some independent continuants c, d and for some time t: b quality_of c at t & b quality_of d at t. (axiom label in BFO2 Reference: [057-001])
140	-----syndrome	OGMS	A pattern of signs and symptoms that typically co-occur.
141	----realizable entity	BFO	A specifically dependent continuant [snap:SpecificallyDependentContinuant] that inheres in continuant [snap:Continuant] entities and are not exhibited in full at every time in which it inheres in an entity or group of entities. The exhibition or actualization of a realizable entity is a particular manifestation, functioning or process that occurs under certain circumstances.
142	-----disposition	BFO	A realizable entity [snap:RealizableEntity] that essentially causes a specific process or transformation in the object [snap:Object] in which it inheres, under specific circumstances and in conjunction with the laws of nature. A general formula for dispositions is: X (object [snap:Object] has the disposition D to (transform, initiate a process) R under conditions C.

143	------disease	OGMS	A disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism.
144	-------acquired genetic disease	OGMS	A disease whose physical basis is an acquired genetic disorder.
145	-------constitutional genetic disease	OGMS	A disease whose physical basis is a constitutional genetic disorder.
146	------emotional action tendencies	EMO(OMD)	NA
147	------homeostasis	OGMS	NA
148	-------abnormal homeostasis	OGMS	Homeostasis that is clinically abnormal for an organism of a given type and age in a given environment.
149	-------normal homeostasis	OGMS	Homeostasis of a type that is not clinically abnormal.
150	------predisposition to disease of type X	OGMS	A disposition in an organism that constitutes an increased risk of the organism's subsequently developing the disease X.
151	-------genetic predisposition to disease of type X	OGMS	A predisposition to disease of type X whose physical basis is a constitutional abnormality in an organism's genome. This abnormality is the physical basis for the increased risk of acquiring the disease X.
152	------function	BFO2	A realizable entity [snap:RealizableEntity] the manifestation of which is an essentially end-directed activity of a continuant [snap:Continuant] entity in virtue of that continuant [snap:Continuant] entity being a specific kind of entity in the kind or kinds of contexts that it is made for.
153	-----role	BFO	A realizable entity [snap:RealizableEntity] the manifestation of which brings about some result or end that is not essential to a continuant [snap:Continuant] in virtue of the kind of thing that it is but that can be served or participated in by that kind of continuant [snap:Continuant] in some kinds of natural, social or institutional contexts.
154	------author role	IAO	A role inhering in a person or organization that is realized when the bearer participates in the work which is the basis of the document, in the writing of the document, and signs it with their name.
155	--independent continuant	BFO	A continuant [snap:Continuant] that is a bearer of quality [snap:Quality] and realizable entity [snap:RealizableEntity] entities, in which other entities inhere and which itself cannot inhere in anything.
156	---immaterial entity	BFO2	NA
157	----continuant fiat boundary	BFO2	b is a continuant fiat boundary = Def. b is an immaterial entity that is of zero, one or two dimensions and does not include a spatial region as part. (axiom label in BFO2 Reference: [029-001])

158	-----zero-dimensional continuant fiat boundary	BFO2	a zero-dimensional continuant fiat boundary is a fiat point whose location is defined in relation to some material entity. (axiom label in BFO2 Reference: [031-001])
159	-----one-dimensional continuant fiat boundary	BFO2	a one-dimensional continuant fiat boundary is a continuous fiat line whose location is defined in relation to some material entity. (axiom label in BFO2 Reference: [032-001])
160	-----two-dimensional continuant fiat boundary	BFO2	a two-dimensional continuant fiat boundary (surface) is a self-connected fiat surface whose location is defined in relation to some material entity. (axiom label in BFO2 Reference: [033-001])
161	----site	BFO2	b is a site means: b is a three-dimensional immaterial entity that is (partially or wholly) bounded by a material entity or it is a three-dimensional immaterial part thereof. (axiom label in BFO2 Reference: [034-002])
162	----spatial region	BFO2	All instances of continuant [snap:Continuant] are spatial entities, that is, they enter in the relation of (spatial) location with spatial region [snap:SpatialRegion] entities. As a particular case, the exact spatial location of a spatial region [snap:SpatialRegion] is this region itself. BFO2 = (A spatial region is a continuant entity that is a continuant_part_of spaceR as defined relative to some frame R. (axiom label in BFO2 Reference: [035-001]))
163	-----zero dimensional region	BFO2	A spatial region [snap:SpatialRegion] with no dimensions. BFO2 = (A zero-dimensional spatial region is a point in space. (axiom label in BFO2 Reference: [037-001]))
164	-----one dimensional region	BFO2	A spatial region [snap:SpatialRegion] with one dimension. BFO2 = (A one-dimensional spatial region is a line or aggregate of lines stretching from one point in space to another. (axiom label in BFO2 Reference: [038-001]))
165	-----two dimensional region	BFO2	A spatial region [snap:SpatialRegion] with two dimensions. BFO2 = (A two-dimensional spatial region is a spatial region that is of two dimensions. (axiom label in BFO2 Reference: [039-001]))
166	-----three dimensional region	BFO2	A spatial region [snap:SpatialRegion] with three dimensions. BFO2 = (A three-dimensional spatial region is a spatial region that is of three dimensions. (axiom label in BFO2 Reference: [040-001]))
167	---material entity	BFO	An independent continuant [snap:IndependentContinuant] that is spatially extended whose identity is independent of that of other entities and can be maintained through time. Note: Material entity [snap:MaterialEntity] subsumes object [snap:Object], fiat object part [snap:FiatObjectPart], and object aggregate [snap:ObjectAggregate], which assume a three level theory of granularity, which is inadequate for some domains, such as biology.
168	----congenital malformation	OGMS	A structurally anomalous part of an organism acquired during fetal development and present at birth (but not necessarily hereditary) which is hypothesized to be harmful for the organism.

169	----disorder	OGMS	A material entity which is clinically abnormal and part of an extended organism. Disorders are the physical basis of disease.
170	-----epigenetic disorder	OGMS	A disorder whose etiology involves (1) a modification to the patient's genomic DNA which leads to alterations in the normal expression pattern of the genome, but is (2) not a change in the nucleotide sequence.
171	-----genetic disorder	OGMS	A disorder whose etiology involves an abnormality in the nucleotide sequence of an organism's genome.
172	------acquired genetic disorder	OGMS	A genetic disorder acquired by a single cell in an organism that leads to a population of cells within the organism bearing the disorder.
173	------constitutional genetic disorder	OGMS	A genetic disorder inherited during conception that is part of all cells in the organism.
174	----fiat object	BFO2	A material entity [snap:MaterialEntity] that is part of an object [snap:Object] but is not demarcated by any physical discontinuities. Bfo2 = (b is a fiat object part = Def. b is a material entity which is such that for all times t, if b exists at t then there is some object c such that b proper continuant_part of c at t and c is demarcated from the remainder of c by a two-dimensional continuant fiat boundary. (axiom label in BFO2 Reference: [027-004]))
175	----injury	OGMS	A part of an organism that has undergone a change in structural integrity and has a higher chance of dysfunction or causing dysfunction in another structure.
176	----material information bearer	IAO	An information bearer is a material_entity, such as a hard drive, upon which an information content entity generically depends.
177	----object	BFO	A material entity [snap:MaterialEntity] that is spatially extended, maximally self-connected and self-contained (the parts of a substance are not separated from each other by spatial gaps) and possesses an internal unity. The identity of substantial object [snap:Object] entities is independent of that of other entities and can be maintained through time.
178	-----organism	OMD/EMO	[Will Hsu] A living object that has (or can develop) the ability to act or function independently
179	----object aggregate	BFO	A material entity [snap:MaterialEntity] that is a mereological sum of separate object [snap:Object] entities and possesses non-connected boundaries.
180	-----extended organism	OGMS	An object aggregate consisting of an organism and all material entities located within the organism, overlapping the organism, or occupying sites formed in part by the organism.
181	-----organism population	OGMS	An aggregate of organisms of the same type.

182	----pathological anatomical structure	OGMS	An anatomical structure (FMA) is pathological whenever (1) it has come into being as a result of changes in some pre-existing canonical anatomical structure, (2) through processes other than the expression of the normal complement of genes of an organism of the given type, and (3) is predisposed to have health-related consequences for the organism in question manifested by symptoms and signs.
183	----pathological formation	OGMS	NA
184	----photographic print	IAO	A photographic print is a material entity upon which a photograph generically depends.
185	----portion of pathological body substance	OGMS	NA
186	---mental functioning related anatomical structure	OMD/EMO	
187	----geographical location	IAO	A continuant [snap:Continuant] that inheres in or is borne by other entities. Every instance of A requires some specific instance of B which must always be the same.
188	-occurrent	BFO	An entity [bfo:Entity] that has temporal parts and that happens, unfolds or develops through time. Sometimes also called perdurants.
189	--emotional behavioral process	EMO(OMD)	NA
190	--processual entity	BFO	An occurrent [span:Occurrent] that exists in time by occurring or happening, has temporal parts and always involves and depends on some entity.
191	---bodily process	OGMS	NA
192	----mental process	OMD/EMO	NA
193	-----appraisal process	EMO(OMD)	[Will Hsu] the process that the classification of someone or something with respect to its worth
194	----pathological bodily process	OGMS	A bodily process that is clinically abnormal.
195	-----pathological derivation	OGMS	A pathological bodily process in which matter is reorganized in such a way as to give rise to new pathological formations which take the place of entities existing earlier.
196	-----pathological invasion	OGMS	NA
197	-----pathological transformation	OGMS	A pathological bodily process in which a canonical anatomical structure becomes a pathological anatomical structure.
198	----physiological response	EMO(OMD)	NA
199	---clinical history taking	OGMS	An interview in which a clinician elicits a clinical history from a patient or from a third party who is reporting on behalf of the patient.

200	---convalescence	OGMS	A processual entity during which a patient participating in a disease course gradually returns to participating in a canonical life course.
201	---diagnostic process	OGMS	An interpretive process that has as input a clinical picture of a given patient and as output an assertion (diagnostic statement) to the effect that the patient has a disease of such and such a type.
202	---disease course	OGMS	The totality of all processes through which a given disease instance is realized.
203	----acute disease course	OGMS	a disease course with a rapid onset but typical unfolding of signs and symptoms after this rapid onset.
204	----chronic disease course	OGMS	A disease course that (a) does not terminate in a return to normal homeostasis and (b) would, absent intervention, fall within abnormal homeostatic range.
205	----progressive disease course	OGMS	A disease course that (a) does not terminate in a return to normal homeostasis and (b) would, absent intervention, involve an increasing deviation from homeostasis.
206	----transient disease course	OGMS	A disease course that terminates in a return to normal homeostasis.
207	---etiological process	OGMS	A process in an organism that leads to a subsequent disorder.
208	---fiat process part	BFO	A processual entity [span:ProcessualEntity] that is part of a process but that does not have bona fide beginnings and endings corresponding to real discontinuities.
209	---health care process	OGMS	A social process that has at least one human participant and that includes as parts the treatment, diagnosis, or prevention of disease or injury--or the following of instructions of another human for treatment, diagnosis, or prevention--of a participant in the process
210	----health care encounter	OGMS	A temporally-connected health care process that has as participants an organization or person realizing the health care provider role and a person realizing the patient role. The health care provider role and patient are realized during the health care encounter
211	-----ED encounter	OGMS	NA
212	-----hospitalization	OGMS	NA
213	-----inpatient encounter	OGMS	NA
214	-----outpatient encounter	OGMS	NA
215	---inflammation process	OGMS	A process which is a response by an organism's tissues that is generally identified by swelling or localized pain
216	---laboratory test	OGMS	A measurement assay that has as input a patient-derived specimen, and as output a result that represents a quality of the specimen.
217	---life course	OGMS	A processual entity which has as parts all the processes in which a given organism is participant.

218	---physical examination	OGMS	A sequence of acts of observing and measuring qualities of a patient performed by a clinician; measurements may occur with and without elicitation.
219	---planned process	IAO	A processual entity that realizes a plan which is the concretization of a plan specification.
220	----assigning a centrally registered identifier	IAO	a planned process in which a new CRID is created, associated with an entity, and stored in the CRID registry thereby registering it as being associated with some entity
221	----associating information with a CRID in the CRID registry	IAO	A planned process in which a CRID registry associates an information content entity with a CRID symbol
222	----data item extraction from journal article	IAO	a planned process in which journal articles are read or processed and data items are extracted, typically for further analysis or indexing
223	----documenting	IAO	a planned process in which a document is created or added to by including the specified input in it.
224	----investigation	IAO	a planned process that consists of parts: planning, study design execution, documentation and which produce conclusion(s).
225	----looking up a CRID	IAO	A planned process in which a request to a CRID registry is made to return the information associated with a CRID symbol
226	----study design execution	IAO	a planned process that realizes the concretization of a study design
227	---process	BFO	A processual entity [span:ProcessualEntity] that is a maximally connected spatiotemporal whole and has bona fide beginnings and endings corresponding to real discontinuities.
228	----history	BFO2	A history is a process that is the sum of the totality of processes taking place in the spatiotemporal region occupied by a material entity or site, including processes on the surface of the entity or within the cavities to which it serves as host. (axiom label in BFO2 Reference: [138-001])
229	----process profile	BFO2	b is a process_profile =Def. there is some process c such that b process_profile_of c (axiom label in BFO2 Reference: [093-002])
230	---process aggregate	BFO	A processual entity [span:ProcessualEntity] that is a mereological sum of process [span:Process] entities and possesses non-connected boundaries.
231	---process boundary	BFO	A processual entity [span:ProcessualEntity] that is the fiat or bona fide instantaneous temporal process boundary.
232	---processual context	BFO	An occurrent [span:Occurrent] consisting of a characteristic spatial shape inhering in some arrangement of other occurrent [span:Occurrent] entities. Processual context [span:ProcessualContext] entities are characteristically entities at or in which other occurrent [span:Occurrent] entities can be located or occur.

233	---prophylaxis	OGMS	A planned process that has the objective to reduce the risk of acquiring one or more disorders.
234	---treatment	OGMS	A processual entity whose completion is hypothesized (by a healthcare provider) to alleviate the signs and symptoms associated with a disorder
235	--spatiotemporal region	BFO	An occurrent [span:Occurrent] at or in which processual entity [span:ProcessualEntity] entities can be located.
236	---connected spatiotemporal region	BFO	A spatiotemporal region [span:SpatiotemporalRegion] that has temporal and spatial dimensions such that all points within the spatiotemporal region are mediately or immediately connected to all other points within the same spatiotemporal region [span:SpatiotemporalRegion].
237	----spatiotemporal instant	BFO	A connected spatiotemporal region [span:ConnectedSpatiotemporalRegion] at a specific moment.
238	----spatiotemporal interval	BFO	A connected spatiotemporal region [span:ConnectedSpatiotemporalRegion] that endures for more than a single moment of time.
239	---scattered spatiotemporal region	BFO	A spatiotemporal region [span:SpatiotemporalRegion] that has spatial and temporal dimensions and every spatial and temporal point of which is not connected with every other spatial and temporal point of which.
240	--temporal region	BFO	An occurrent [span:Occurrent] that is part of time.
241	---connected temporal region	BFO	A temporal region [span:TemporalRegion] every point of which is mediately or immediately connected with every other point of which.
242	----temporal instant	BFO	A connected temporal region [span:ConnectedTemporalRegion] of a single moment of time.
243	----temporal interval	BFO	A connected temporal region [span:ConnectedTemporalRegion] lasting for more than a single moment of time.
244	---scattered temporal region	BFO	A temporal region [span:TemporalRegion] every point of which is not mediately or immediately connected with every other point of which.
245	-signs	OGMS	A quality of a patient, a material entity that is part of a patient, or a processual entity that a patient participates in, any one of which is observed in a physical examination and is deemed by the clinician to be of clinical significance.
246	--vital signs	OGMS	A physical sign in which a non-zero value is standardly considered to be an indication that the organism is alive.
247	-symptom	OGMS	A quality of a patient that is observed by the patient or a processual entity experienced by the patient, either of which is hypothesized by the patient to be a realization of a disease.
248	--pain	OGMS	NA

15 Generating Self-Explanatory Data Repositories from Clinical Research Datasets using Referent Tracking[3]

15.1 Introduction

One common application of ontologies is to the integration of information residing in heterogeneous data collections. The assumption is that there are queries that can be resolved when run over the combined data that would remain unanswered if addressed only to its constituent datasets taken singly [111]. Different paradigms for such integration have been proposed, including *mediation* [112], *federation* [113], *data warehousing* [114], and the *Ontology-Based Data Access* (OBDA) methods described in [115]. It is well recognized that such paradigms, to be effective, require some form of ontology-based mapping between the schemas of the separate databases of a sort that takes account of the semantics not only of the data but also of the data types by means of which these data are stored [116]. On the OBDA approach, which draws on the Semantic Database paradigm initiated already in the 1980s, the data sources and what is called the conceptual layer of the information system are kept separate and independent. But OBDA goes further and argues that if information integration is to work then much more detail is required of the mappings than is commonly supplied. Specifically, it argues that attention must be paid not merely to the T (for 'terminology') Box of statements about *general* terms (or classes in the ontology) but also to the A (for 'assertion') Box of instance data. OBDA thus requires mechanisms not merely for mapping data *fields* to classes in the ontologies but also for mapping individual data *values* to corresponding instances of these classes. This requires specifying how instance identifiers can be composed or resolved starting out from data values in a way that allows the construction of an ABox suitable for answering queries about instances [117].

The latter is, we believe, a critical issue not least in the clinical context, in virtue of the fact that, as users of clinical record systems are only too well aware, data values do not always denote what is suggested by the variable or fieldname under which they appear. Suppose, for example, that in a patient's record for the variable *phenotypic gender* it is not the literal values 'male' and 'female' which are found, but rather the coded values '0' and '1', respectively. If '0' or '1' appear in the corresponding field of the dataset for a given patient, then it is reasonably safe to create an ABox statement to the effect that this patient's phenotypic gender is indeed an instance of the corresponding ontology class. If, on the other hand, no data value is found, then it should not be assumed that the given patient does not have a phenotypic gender. And if, on yet another hand, a value of '9' – documented as signifying 'unknown' – is found, then this should not lead to an ABox assertion to the effect that the patient in question has a phenotypic gender that is an instance of some third kind (*unknown gender*) that is neither *male* nor *female*. An *absent finger*, similarly, is not a special kind of finger; and an *adenoidectomy without tonsillectomy* is not a special kind of *removal of the adenoids*. [90] Sadly, the use of coded values often involves not only (a) employment of terms (such as '0' and '1') in ways which depart from their standard meanings but also (b) illegitimate hypostatization of associated classes (types).

Clinical record systems also reveal a surprising amount of hidden and ambiguous data. *Hidden* data is illustrated by an assertion such as 'The patient's strength of right foot plantar flexion is

[3] Derived material from this section has been accepted for publication as: Ceusters W, Hsu CY, Smith B. Generating Self-Explanatory Data Repositories from Clinical Research Datasets using Referent Tracking. International Conference on Biomedical Ontologies, ICBO 2014, Houston, Texas, Oct 6-9, 2014.

3/5', which is elliptical for: 'The *measurement* of the strength of the patient's right foot plantar flexion yielded a value of 3 on a scale from 0 to 5.' *Ambiguous* data is illustrated by a case in which a doctor refers to a 'patient's father' but fails to take account of the fact that the patient in question has both a biological and an adoptive father. An additional problem arises because the information required to create mappings from clinical research data values to ontologies is often scattered over multiple data dictionaries and guidelines made available when processing data collected on the basis of multiple standardized questionnaires and case report forms.

The goal of the work reported below was to determine (1) what kinds of ambiguous, hidden and inappropriately coded information are encountered in such data collections, and (2) whether it is possible to provide a complete and explicit representation of clinical research datasets in a way that takes account of the sorts of constraints and provisions typically documented in data dictionaries and other data-related sources of guidance. The hypothesis is that, even where it is not possible to provide a completely accurate representation of the part of reality described by a given body of data, identifying the degree to which data are affected by problems of the sorts mentioned would itself yield a useful resource for the measurement of data quality and provide a means of avoiding similar problems in future clinical studies – for example through better training of questionnaire and form designers and of data providers.

15.2 Foundations

15.2.1 Ontological Realism

Our work is based on two prior developments: ontological realism and referent tracking.

The advocate of ontological realism holds that the most effective way to ensure mutual consistency of ontologies over time, and to ensure that ontologies are maintained in such a way as to keep pace with advances in empirical research, is to view terms in ontologies as representations of the types or universals in reality that are represented by the general terms (such as 'cell' or 'lung' or 'inflammation') used in the scientific literature [82]. Only ontologies of this sort, we believe, can justify the sort of investment needed to effectuate ontology-based data integration, since only ontologies of this sort will provide an environment that is sufficiently robust (which means: able to maintain stability of content even while taking account of scientific advance) to support the creation of the sorts of stable mappings useful to integration.

Recent inquiries suggest that ontological realism is usurping the older paradigm of 'Medical Knowledge Representation' formally dominant in the clinical domain [99]. Under the latter, ontologies and terminologies are conceived as representations not of what is general in reality, but rather of what are called 'concepts' (conceived, roughly, as ideas in people's heads). Concept-based terminologies will of course remain of importance to clinical informatics in the future, not least because of the massive quantities of clinical data already collected and described in their terms, and for this reason they play a role also in the work described below. To serve as inputs to effective integration, however, they require a basis in realist ontologies along the lines described.

15.2.2 Referent Tracking

The principal thrust of our efforts here, accordingly, is to attempt to raise the quality of the instance data available for information-driven clinical science by providing an explicit representation of the particulars which are the instances of the classes in realist ontologies. It is to this end that the methodology of referent tracking (RT) has been developed, beginning already with our [77], in which we presented an algorithm to detect the sorts of ambiguous and

implicit information that are typically to be found in today's highly structured electronic health records (EHRs). Briefly, we showed how data from one specific EHR application needs to be decomposed to provide adequate representations of particulars in reality. At the same time we described the ontological principles on which such decomposition needs to be based in order to allow similar efforts to be extended to other EHR applications by parties interested in data integration and consistency.

In [75] we showed how to build data repositories whose content can be expressed as a collection of what we called *Referent Tracking Tuples* (RTT). An RTT is (in the sorts of cases of importance to us here) an assertion about a particular, i.e. an entity in reality that exists or occurs in some region of space and time [118], that is expressed by means of one or more Instance Unique Identifiers (IUI). Each RTT follows a semi-formal syntax that is close to the one used for instance-level relationships in the definitions of the Relation Ontology [119]. Thus it rests on a distinction between continuants, that is entities (such as molecules, cells, organisms) that exist through time and undergo changes; and occurrents, which are the processes in which continuants participate, for instance when some change occurs 0.

Consider, first, a simple assertion about some relationship in which some continuant is involved (for instance to the effect that John is *part_of* this group of doctors, or that Mary *participates_in* that clinical encounter). Leaving aside certain house-keeping parameters not of relevance to our concerns here, the corresponding RTT will be of the form *'x p-rel y t-rel t'*, where:

- '*x*' is a singular and (internally) globally unique instance identifier (IUI) denoting the particular described,
- '*y*' is either (1) an IUI denoting another particular or (2) a representational unit drawn from either a realism-based ontology or a concept-based terminology,
- '*p-rel*' expresses a relationship obtaining between *x* and *y*,
- '*t*' denotes a particular temporal region,
- '*t-rel*' expresses the relationship obtaining between the temporal region denoted by '*t*' and the exact, i.e. maximal, temporal region during which *p-rel* obtains between *x* and *y*. (Note that in [119] '*t-rel t*' is restricted to '*at t*' in the meaning of '*obtains at least during t, perhaps also at other times*'; in the RTT framework, in contrast, it can represent any of the temporal relationships defined in [120].)

For assertions that do not mention a continuant, the corresponding RTT is of the simpler form '*x p-rel y*', where '*x*', '*p-rel*' and '*y*' carry the same meaning as in the above. An RDF implementation of both forms of tuples is described in [76]. Note that an IUI in our terminology is not analogous to what philosophers call a 'definite description' [121]. Rather, it is an artifact (a meaningless alphanumeric string) that is designed to serve as a pure denotator. Typically, an IUI is assigned to its referent in a process involving both direct human ostension and software, as when a clinician tells his information system to assign an IUI to the fracture in the leg of the patient he is now examining. When we refer to an IUI as 'globally unique' then what we mean is that it has been generated by a piece of random-string-generating software that has been incorporated into some referent tracking system (RTS) in a way that ensures (with a very, very, very high degree of probability) that the string in question is not used to designate some other entity at some other location within the jurisdiction in which this RTS is applied [118]. IUIs fulfil similar goals as URIs and in semantic web applications of the RT paradigm IUIs will be concretized as URIs.

15.2.3 Correcting implicit and ambiguous information through referent tracking

One goal of the RT methodology is to convert ambiguous into non-ambiguous clinical data by means of IUIs. Consider the problem which arises when at t_1 a clinician refers in his patient's EHR to *some* instance *of a given type* without specifying *which* instance is intended, as in an assertion such as '*John has a benign duodenal polyp*'. The problem arises because, to capture such assertions, EHR technology typically employs merely a general diagnostic code drawn from some terminology or ontology. As a consequence, when at a later time t_2 an entry is made in John's EHR to the effect that he has a malignant (rather than a benign) duodenal polyp, then the system does not allow clinically important inferences to be drawn to the effect that it is either (a) the very same polyp that has turned malignant or (b) some second polyp that has arisen at some other location in John's duodenum [75].

We can now represent the earlier situation using the following RTTs:

- #1 **part-of** #2 **at** t_1 (1)
- #1 **instance-of** BENIGN DUODENAL POLYP **at** t_1 (2)

where '**#**1' and '**#**2' are IUIs denoting the polyp and John, respectively. Following the principles of ontological realism, 'benign duodenal polyp' and 'malignant duodenal polyp' must be representational units from a realism-based ontology, and 'instance-of' must be defined as in the Relation Ontology [119]. (If on the other hand we are willing to take over the terms in question from a concept-based terminology, then 'instance-of' would be replaced by the 'particular-to-concept-relation' (PtoCo) defined in [76].)

To see how the mentioned ambiguity disappears on the RT approach we note that the first of the two distinguished scenarios would be captured by (1) and (2) together with an additional RTT of the form:

- #1 **instance-of** malignant duodenal polyp **at** t2 (3)

The alternative scenario, in contrast, in which a second, malignant, polyp came into existence at a time subsequent to the appearance of the first, would be represented by (1) and (2) together with RTTs employing a new IUI for the second polyp, as follows:

- #3 **part-of** #2 **at** t_2 (4)
- #3 **instance-of** MALIGNANT DUODENAL POLYP **at** t_2. (5)

This methodology for disambiguation is most effective when its principles are applied at the time of data collection and registration. But it brings benefits also when used in *post hoc* translations. [122] There, too, it will help us to make explicit all implicit assumptions that need to be taken into account in order to interpret the data correctly, some of which result from defective information models or flawed practices – for example registering ICD-9-CM code 659.7 – '*Abnormality in fetal heart rate or rhythm*' in the diagnosis field of a mother's EHR rather than in the EHR of her fetus.

15.3 Materials and methods

15.3.1 Principles

Datasets were made available to us as spreadsheet tables (henceforth called 'source tables'). Each row in the body of each such source table is a collection of data items obtained from a single patient. Each column is a collection of data items resulting from some specific type of observation. If a header row is present, then the terms in its cells indicate what sorts of observation results are reported in the columns beneath.

The work reported on here involved the following steps for each dataset completely available at the time of the effort:

1. cross-check each variable in the study set with the variable codebook and technical report for appropriate coding values, field names, and field descriptions, and categorize the various sorts of data and metadata involved, for instance whether a variable must have a value for each patient;
2. build an executable template that specifies, for each of the possible data values, how that value's referent must be analyzed in RT terms, thereby applying the following data expansion algorithm (taken from [77]):
   2.1. identify all the possible particulars that are explicitly referred to by a specific data value when applied to a specific patient,
   2.2. determine for each particular identified in (2.1) whether it is a dependent or an independent entity [82],
   2.3. if a particular is an ontologically or existentially dependent *continuant*, identify the independent continuant on which it depends; if a particular is an *occurrent*, identify the continuants that participate in it,
   2.4. repeat steps 2.2 and 2.3 as required for the entities identified,

Table 1: 16 sample lines from the data dictionary portion of the template for data expansion of the variables ('Var') 'id', 'sex' and 'q3' used in the original dataset.

L	Var	DT	REF	Min	Max	Val
1		IM	patient_study_record			
2	Id	LV	patient_identifier			
3	Id	IM	patient			
4	sex	CV	gender			
5	sex	CV	male			0
6	sex	CV	female			1
7	sex	UA	sex	BLANK	BLANK	
8	q3	CV	no_pain_in_ lower_face			0
9	q3	CV	pain_in_ lower_face			1
10	q3	IM	in_the_past_month			
11	q3	IM	lower_face			
12	q3	IM	time_of_q3_concretization			
13	q3	RP	an_8_gcps_1	0	0	0
14	q3	UP	an_8_gcps_1	1	10	0
15	q3	UA	an_8_gcps_1	BLANK	BLANK	1
16	q3	JA	an_8_gcps_1	BLANK	BLANK	0

Legend: 'L' = Line number in this table, 'Var' = Variable, 'DT' = Data Type (with possible values being 'LV' = Literal Value, 'CV' = Coded Value, 'UA' = Unjustified Absence, 'IM' = IMplicit reference, 'RP' = Redundant Presence (RP), 'UA' = Unjustified Absence, 'JA' = Justified Absence), 'REF' = Reference, 'Min' = lowest possible value for variable, 'Max' = highest possible value for variable, 'Val' = possible value for variable.

3. select from realism-based ontologies the representational units that denote universals (or defined classes in the sense of 0), whose instances are directly referred to in the dataset or are discovered through steps 2.1–2.4,
4. implement an algorithm that uses the template developed in step 2 to generate for each patient described in the dataset the collection of RTTs that provides a realism-based representation of his situation.

Tables 1 and 2 should be viewed together (with the former on the left and the latter on the right). The whole illustrates 16 sample lines from the template, which have been somewhat simplified for the purpose of this paper. The template has been created for a specific dataset and is applied to all patients therein by the software mentioned in step 3.

Table 2: Ontology portion of the template for data expansion for the 16 sample lines of Table 1.

L	IUI(I)	IUI(P)	P-Type	P-Rel	P-Targ	Trel	Time
1		#psrec-	DATASET-RECORD			At	T
2	#pidL-	#pid-	DENOTATOR	**denotes**	#pat-	At	T
3	#patL-	#pat-	PATIENT			At	T
4	#patgL-	#patg-	GENDER	**inheres-in**	#pat-	At	T
5		#patg-	MALE-GENDER	**inheres-in**	#pat-	At	T
6		#patg-	FEMALE-GENDER	**inheres-in**	#pat-	At	t
7		#patgL-	UNDERSPECIFIED-ICE			At	t
8	#q3L0-	#pat-		**lacks-participant**	PAIN	At	#tq3-
9	#q3L1-	#pq3-	PAIN	**participant**	#pat-	At	#tq3-
10		#tq3-	MONTH-PERIOD				
11		#patlf-	LOWER-FACE	**part-of**	#pat-	At	t
12		#cq3-	TIME-PERIOD	**after**	#tq3-		
13	#q3L-	#q3L-		**corresponds-with**	#q3L0-	At	t
14	#q3L-	#q3L-	DISINFORMATION			At	t
15	#q3L-	#q3L-	UNDERSPECIFIED-ICE			At	t
16	#q3L-	#q3L-	JUSTIFIED-BLANK-ICE			At	t

Legend: 'L' = Line number in this table, 'IUI(I)' = prefix for generating an IUI proxy for the information content entity which refers to the corresponding value for the variable under 'Var' (Table 1) for the patient being processed, IUI(P) = prefix for generating an IUI proxy for whatever is denoted by this information content entity, P-Type = ontological type of the entities denoted by instantiated IUI(P)s, P-Rel = relation between the entity denoted by an instantiated IUI(P) and the entity denoted by an instantiated P-Targ, 'Trel' = temporal relation, 'Time' = temporal period during which P-rel holds. Only entries relevant to the discussion in this paper are shown.

Together these tables present the two parts of the template manually produced for the variables 'id', 'sex' and 'q3' following steps 1–3 as described above for the German dataset (section 3.1) which was taken as example. Table 1 contains a representation of metadata derived from the data dictionary and other documentation provided by the original compilers of the dataset. Table 2 contains information about how specific values for the variables need to be interpreted and how the portion of reality from which they are derived should be represented using RTTs. Each line in Table 1 specifies conditions which, when satisfied, lead to the generation of RTTs based on the RTT templates listed in the corresponding line of Table 2. In the sequel, we explain how lines 2 to 16 as displayed in these tables have been constructed. Line 1 in each table is just that part of the template that, on execution, will assign an IUI-proxy – here generically referred to as #ps-rec (for 'patient study record') – to the entire record of the patient in the study set. Such proxies stand in for corresponding IUIs which will figure in the expanded dataset stored in an RTS.

15.3.1.1 Data categorization and expansion

The approach for building the template proposed here is based on a (to us) obvious distinction between data, on the one hand, and what these data are about, on the other. It thus provides not only for RTTs which describe in an explicit way those portions of reality which involve the particulars described by data items in a dataset but also for RTTs which describe the portions of reality comprising these data items themselves. Some RTTs, for example those using the relation *aboutness*, capture both.

This combination allows us to describe particulars that are only implicitly referred to in the dataset and to provide information about what we shall identify below as 'correspondences' between different data items in a single dataset. It also allows us to assert which data items are unjustifiably or redundantly present or absent in a dataset, and so forth.

By following steps 1 to 2.4 from section **2.** above, our data analyst concluded that the variable 'id' contains literal values (noted as 'LV' under Data Type ('DT') in line 2) of which each value functions as a patient identifier (under 'REF') within the dataset. The analyst concluded further (step 2.3) that for a patient identifier to be appropriately assigned, there has to be a patient, which in turn led to the creation of line 3 in Table 1, stating that the existence of a patient is implicit ('IM' under 'DT') given the existence of a patient identifier.

The variable 'sex' (used in the German dataset), in contrast, was found to contain coded values ('CV') for a subject's gender, possible values as listed in the 'Val' column being '0' for male and '1' for female, thus leading to lines 5 and 6. ('Coded value' means that these are not the same '0' and '1' as would be used, for example, when recording a number of offspring.) The analyst further asserted that all patients have a gender (line 4), and that a corresponding value must be registered in the dataset (line 7). The latter is expressed by the presence of 'UA' – Unjustified Absence – in the 'DT'-column and, by convention by the presence of 'BLANK' in both the minimal and maximal allowed values (see section *5.3* below).

Table 3 provides some statistics concerning the lines from out of which the data translation template for the study set is composed, and on the extent to which each of these lines were in fact applied to the patient population described in the study set. The table shows, for instance, that unjustified absences and presences were encountered, albeit in a small percentage of cases, and that on average for each variable and for each patient roughly 3 implicit particulars needed to be accounted for. It shows that the increase in the size of the dataset resulting from applying this methodology is, for the German dataset, roughly 300%, and also that the quality of this dataset (measured in terms of UA, RP and UP) is quite good.

Table 3. Minimal, maximal and average occurrence of lines in Table 1 and 2 (1) constructed in the template per variable (n=161) in the study set (left block), and (2) actually applied per patient (n=390) (right block).

	Template			**Patients**		
	Av. (SD)	**Min**	**Max**	**Av. (SD)**	**Min**	**Max**
CV	3.57 (2.27)	0	11	0.82 (0.38)	0	1
IM	2.79 (1.43)	0	6	2.69 (1.46)	0	6
UA	0.16 (1.02)	0	12	0.01 (0.09)	0	10
JA	0.16 (1.02)	0	12	0.04 (0.34)	0	12
RP	0.13 (0.98)	0	12	0.01 (0.10)	0	11
UP	0.13 (0.98)	0	12	0.00 (0.01)	0	5

Legend: 'Av.' – average, 'SD' – standard deviation.

15.3.1.2 Referent tracking tuple patterns

Table 2 depicts the ontological part of the template that the analyst builds as input for the algorithm developed in step 3 from section **2** above. Each line in Table 2 contains what we will call a 'referent tracking tuple pattern' (RTTP). The conditions to execute the pattern associated with a given line as displayed in Table 1 are asserted in the line with the same number in Table 2. The algorithm iterates over the dataset moving from one patient to the next and checking, in relation to each successive patient's values, for a match with the conditions in the template in Table 1. Any match leads the algorithm to generate an RTT using the corresponding RTTP as basis and describing for each patient either the relevant particulars referenced in the dataset or the relevant elements of the dataset itself.

Each particular that needs to be represented in an RTT must be denoted by some IUI that is created in a referent tracking system [118], for example as part of a clinical encounter. The algorithm discussed here thus does not create these IUIs themselves; rather it creates what we call 'proxies' for these IUIs, which serve as unique identifiers within the context of the self-explanatory datasets created from the original datasets by means of the template. The algorithm creates these proxies by concatenating the entries under 'IUI(I)' or 'IUI(P)' on the corresponding lines in Table 2 with the value of 'Id' for the patient being processed. For example, for line 4 and for patient 2053 it generates: '#patgL-2053' and '#patg-2053', respectively.

We will first illustrate how the approach works using line 4 as our example (using 'Ln' as abbreviation for 'line n'). We will then proceed in section 5 to a more systematic description of the different sorts of cases. In addressing L4, the data analyst forces the algorithm to an unconditional execution of the RTTP in Table 2 by not providing any entries in the columns 'Min', 'Max' and 'Val' of Table 1. Through the entry '#patgL-' under 'IUI(I)', the analyst forces the algorithm to create, in accordance with the principles of the Information Artifact Ontology [84], a proxy for the IUI denoting that information content entity which expresses the gender of the patient being processed. With the entry '#patg-' under 'IUI(P)', a proxy IUI is created to denote this gender itself.

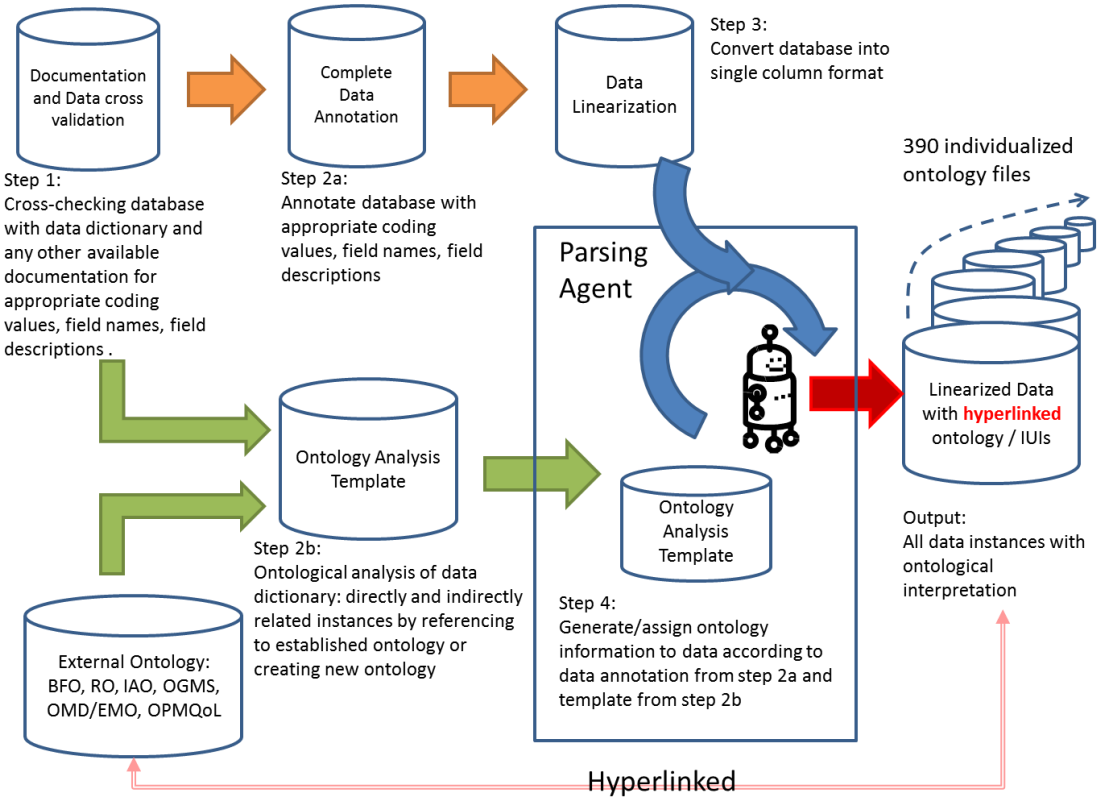
Entries under 'P-Type' are used to express instantiation. Thus when L4 is processed for the patient with id '2053', the particular denoted by #patg-2053 will be translated in RT terms as being an instance of *gender* (we assume for the sake of argument that *gender* qualifies as a universal and that someone's gender is a particular dependent continuant quality; we here ignore the small changes necessary if gender is instead treated as a defined class in the sense of [123]).

Whereas the columns 'IUI(P)' and 'P-Type' form together the template for RTTs expressing instantiation, the columns 'P-Rel', 'P-Targ', 'Trel' and 'Time' form together with column 'IUI(P)' the template for RTTs expressing any other relation. Entries under 'P-rel' contain the name for the relation in question between the entity denoted by an instantiated IUI(P) and the entity denoted by an instantiated P-Targ, 'Trel' indicates a temporal relation, while 'Time' denotes the temporal period during which P-rel holds.

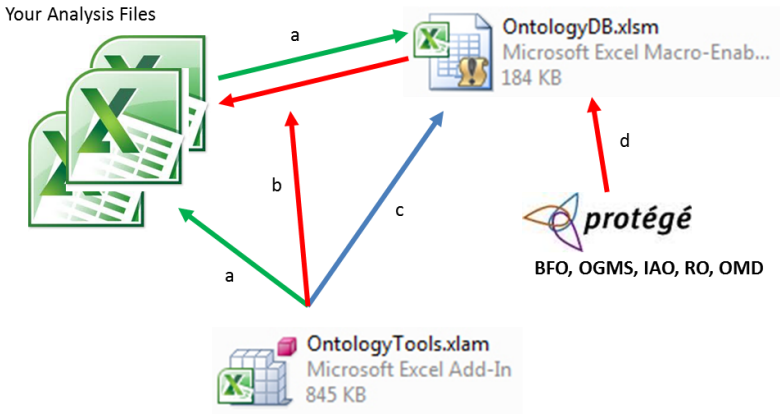
For each line in the template, there must be an entry for either 'P-Rel' or 'P-Type', or both. Whenever, in a given line, the particular denoted by an instantiated IUI(P) or P-Targ is a continuant, there must be an entry for 'Trel' and 'Time' asserting when the relationship P-Rel holds. (We will ignore in what follows both the underspecification of the time-related information in our examples, and certain additional details required by syntactically and semantically correct RTTs [118].)

15.3.2 Implementation

15.3.2.1 Database Ontology Processing Flowchart



Ontology Tools and Ontology DBs are designed to facilitate the process of building the dataset application ontologies. All the steps in the building process can be completed in Microsoft Excel. There are 3 components or files involved in the process: analysis files, OntologyDB and Ontology Tools.



15.3.2.2 Stepwise breakdown of Dataset Ontology Analysis

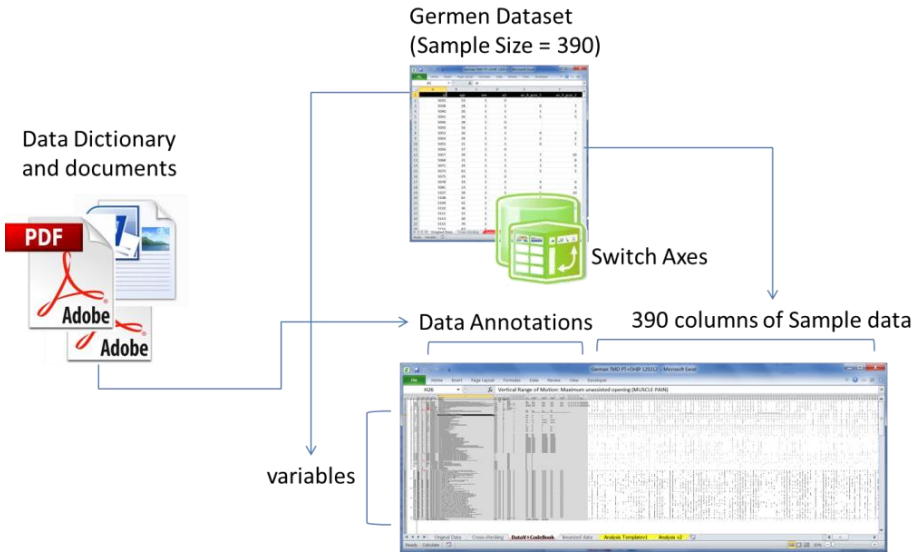
15.3.2.2.1 Step 1: Cross-checking database with data dictionary/documents

None of the datasets were self-explanatory which required considerable cross-checking with the supporting documents and subsequent annotation of variables and values. For the German dataset, for instance, there were 158 heterogeneous variables corresponding to:

- Demographics: id, age, gender
- RDC Pain history: 8 questions + 5 calculations
- RDC Examination: 66 questions (14 different types of data coding ranges)
- Jaw Problem History: 12 questions
- RDC Examination Criteria: 14 criteria determined by calculations
- Oral Health Impact profile Quality of life: 49 Questions + 1 total score

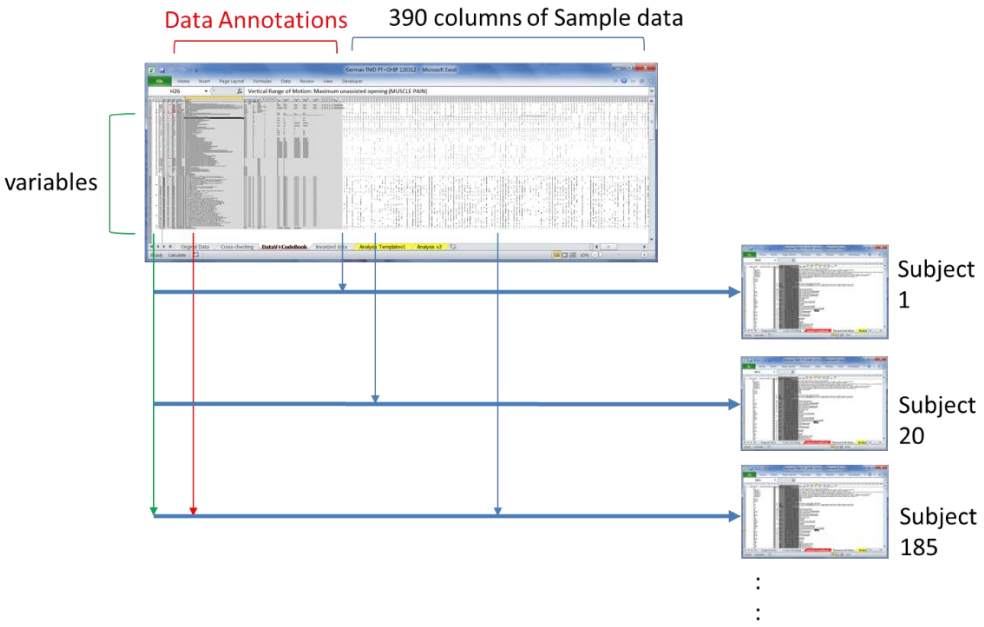
15.3.2.2.2 Step 2a: Annotating the dataset

While confirming the integrity of the data, it is essential to integrate the data dictionary with the dataset. The data annotation step reduces effort and time for future information research. A consistently annotated database will also increase the value of the dataset for reusability

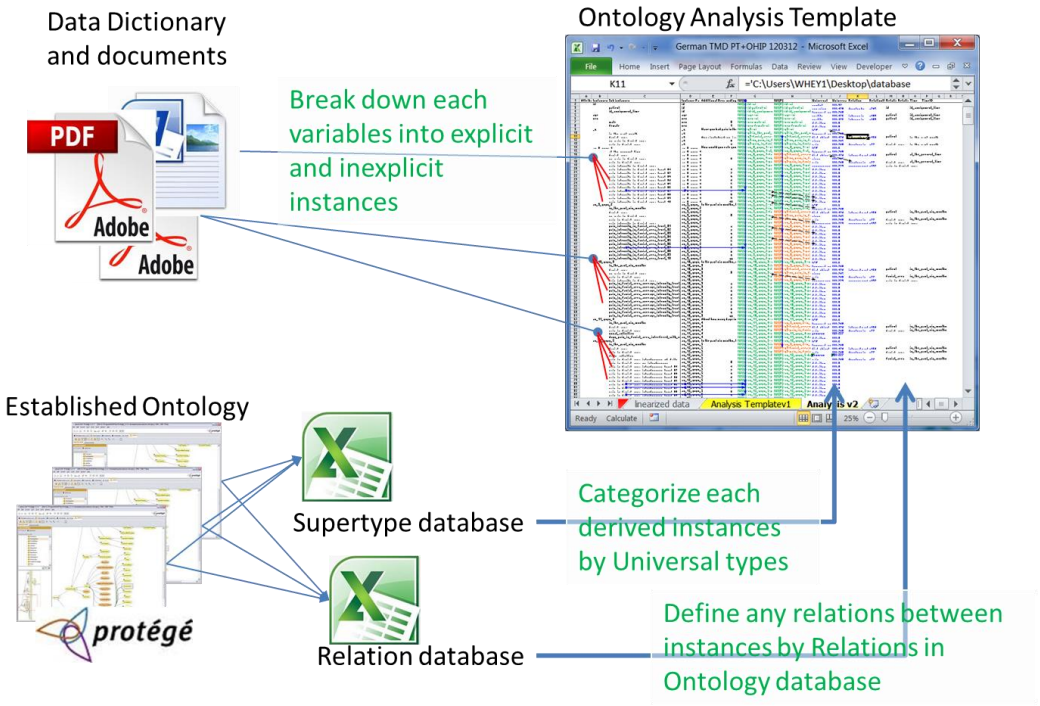


15.3.2.2.3 Step 3: Linearize Data –

This step is to generate a compatible format for the Parsing Agent in order to generate ontologically analyzed data and is performed by an algorithm incorporated inside the OntologyTools/DBtools/DataLinearization.

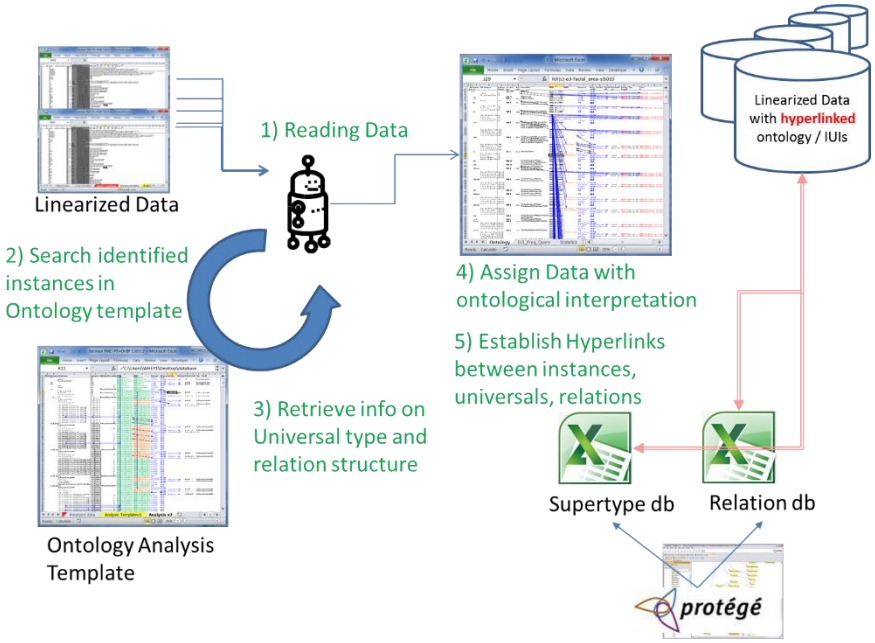


15.3.2.2.4 Step 2b: Ontology Analysis Template



15.3.2.2.5 Step 4: Parsing Agent

This agent implements an algorithm that analyzes each patient's data ontologically according to the ontology template and produces an individual data profile with detailed ontological interpretations.



15.3.2.3 Computing resource assessment

Experiments were carried out to assess the feasibility for this type of processing – i.e. the fully automatic parts of the methodology – on personal desktops or laptops. The following table summarizes the results for the Swedish dataset.

Algorithms	Note (result location)	Processing time (based on 8-core processor)
1 Build Data Bridge	This process combines Data and Data Dictionary into a single worksheet. (output: Data_Bridge)	1-3 minutes
2 Discover Interactions	This algorithm conducts exhaustive searches on consistent patterns between 2 variables for all possible 2-variable combinations (n=204 x 203 = 41412). (output: Discovered_Interactions)	6-10 minutes
3 Generate Ontology Template	The process generates the Ontology Template by 1) converting data descriptions into analyzable tokens, 2) listing all combinations of coding values, 3) inserting all existing variable interactions for each specific variable from step 2 (“Discover Interaction”), 4) generating an accessible data navigation index. (output:AnalysisIndex, Analysis)	5-8 minutes
4 Parser – single subject	The process generates a customized ontology analysis for the specified subject ID and quantify the data pattern (output: ParsedIndex, ParsedData)	30 seconds – 1 minute
5 Parser – continuous	The process analyzes a specified range of subject IDs (output: Parsing Summary)	12 seconds per subject

15.4 Results

15.4.1 Generating self-explanatory representations: applying the templates

Our vision is that data repositories should be maximally explicit and self-explanatory. By 'maximally explicit', we mean that each repository should contain explicit reference to any and all entities in reality that must exist for an assertion encoded in the repository to be a faithful representation of the corresponding part of reality, including the relationships in which these entities stand to each other. By 'self-explanatory' we mean that the data in the repository should be presented in such a way that a researcher seeking to query these data does not need to worry about any idiosyncrasies of the separate datasets that were combined to build the repository. This can be achieved only where the datasets submitted for inclusion in a combined repository have themselves been rendered maximally explicit and self-explanatory in the way described. We shall now demonstrate how, using a template of a sort displayed in Tables 1 and 2 was able to be transformed into one that is maximally explicit and self-explanatory by means of referent tracking tuples.

15.4.1.1 Explicit data items

The study set contains data items about particulars on the side of the patient. Thus they are about a patient's gender, the facial pains he or she has experienced, the clicking noises he or she has heard when opening his or her mouth, and so forth. Ontological realism tells us that each such particular is an instance of at least one universal or type [82]. What the relevant types are is represented in the study set – but typically only very indirectly.

Imagine, now, that the algorithm of step 3 (from section **2** above) is implemented in some software and that this software is processing our study set using the template displayed in Table 2. L3 of the template will cause the assignment of the IUI-proxy *#pat-1* to the patient in the dataset referred to by means of the identifier '1' (where we are assuming that '1' in the dataset is the value for the variable 'id' for that patient). L4 will lead to the assignment of *#patg-1* to this patient's gender. The analyst did not specify any conditions for L4 to be executed, so it is executed unconditionally – reflecting the reality that every patient has a gender, whatever that gender might be. If the value of 'sex' for that patient in the dataset is '0', then L5 will be executed as well. If that value is '1', then not L5 but L6 will be executed. For any other value or no value at all, neither L5 or L6 would be executed. The following collection of assertions would then be generated on the basis of the assumption that the value for 'sex' (the term used for 'gender' in the German dataset).is '0', and this, if the study set itself is faithful to reality, would constitute a faithful RT-representation of the corresponding portion of reality.

- *#pat-1* **instance-of** PATIENT **at** *t* (6)
- *#patg-1* **instance-of** MALE-GENDER **at** *t* (7)
- *#patg-1* **inheres-in** *#pat-1* **at** *t* (8)

(where *italics* is used for particulars, SMALL CAPS for universals, and **bold** for relations involving particulars).

Of course, the study set itself is a particular, and so are the rows and data items that form its parts. More precisely (again following IAO [84]), the study set and its parts are particular information content entities (ICEs). The analyst acknowledged this for instance in L1, which is what led to the assignment of *#psrec-1* to the ICE which denotes the record for the relevant patient in the study set, and *#patgL-1* to the ICE which denotes the gender of patient *#pat-1* (see L4). Since referent tracking implementations also assign IUIs to RTTs, the IUI-proxy *#RTT-*

patg-1-L5 would be assigned to the ICE of which assertion (7) above is a concretization. This means that assertions such as the following can now be added to our RTT repository:

- *#patgL-1* **component-of** *#psrec-1* **at** *t* (9)
- *#RTT-patg-1-L5* **instance-of** RTT **at** *t* (10)
- *#patgL-1* **corresponds-with** *#RTT-patg-1-L5* **at** *t* (11)
- *#patgL-1* **instance-of** DATA-ITEM **at** *t* (12)
- *#patgL-1* **is-about** *#patg-1* **at** *t* (13)
- *#psrec-1* **instance-of** DATASET-RECORD **at** *t* (14)

In summary: assertions of the sorts (9), (12) and (13) are generated on the basis of values encountered in the dataset for all IUI(I)-IUI(P) co-occurrences in the template. Assertions of the sorts (10) and (11) are generated for all lines in which an RTT template is specified and for which the conditions specified in Table 1 are satisfied. And assertion (14) is generated because of L1.

The **corresponds-with** relationship holds between two ICEs (each RTT is an ICE in its own right) whenever they faithfully describe or denote the same portion of reality. Assertions (6) to (8) describe certain portions of the reality on the side of the patient, and (9) to (12) describe information content entities that have some aboutness relation with these portions of reality. (13) then provides the link between the reality on the side of the patient and a description thereof.

15.4.1.2 Referencing implicit information: an example from the study of pain

The variable ‘q3’ in the study set holds responses to the question: *Have you had pain in the face, jaw, temple, in front of the ear or in the ear in the past month?* A positive answer is encoded as ‘1’ (L9), a negative one as ‘0’ (L8). We note that some particulars on the side of the patient to whom the question is addressed (his jaw, temple, past month, etc.) are explicitly referred to in the question yet none of them are referred to in either of the two possible responses. To achieve our objective, explicit reference to some of these particulars has to be created, and this is achieved by means of IM-lines, all of which have under ‘REF’ a textual reference to an entity in reality – or to some configuration of such entities [118] – that must exist for the corresponding ‘Var’ to make sense.

When the template is applied to *#pat-1*, a negative answer to question q3 (L8) would generate an RTT to the effect that the patient lacks participation in an instance of pain – *pain is a process* [97] – by using the lacks-family of relations for negative findings [10]. In case of a positive answer, an IUI for the corresponding pain instance is generated and participation of the patient therein is asserted (*the patient suffers a pain process*). Both answers generate IUIs and corresponding assertions for the patient’s lower face, the time when the question was asked, and the salient period (of one month prior to asking). Note that all of these entities exist whatever the answer supplied by the patient.

15.4.1.3 (Un)justified presence and absence

Template lines with types UA, UP, RP, and JA serve to make explicit the fact that data are or are not missing from a dataset, or that there are data that should not be there. L7, for instance, brings it about that, when no value for the variable ‘sex’ is provided for patient *#pat-1* in the study set, then this is expressed by the appearance in the template of ‘BLANK’ under both ‘Min’ and ‘Max’ (see **3** above). An RTT will then be generated that declares the data item *#patgL-1* to

be an instance of an underspecified ICE. The latter means not that the data item in question is absent, but rather that some information is missing.

An absence or presence of a value for some variable may be justified or unjustified depending on the value of some other variable. The last four lines in Table 1, for example, describe dependencies between the variables 'q3' and 'an_8_gcps_1', the latter containing answers to the question: *How would you rate your facial pain on a 0 to 10 scale at the present time, that is right now, where 0 is "no pain" and 10 is "pain as bad as could be"?* L13 states that when the values for 'q3' and 'an_8_gcps_1' are both '0', then the two ICEs involved correspond-to the same portion of reality. L16 asserts that, if a record in the dataset has a '0' value for the variable q3, and there is no value for the variable 'an_8_gcps_1', then the absence of a value for 'an_8_gcps_1' is justified. This is then documented by means of an RTT that asserts the corresponding ICE to be justifiably blank (concretized by, for instance, an empty cell in that part of the spreadsheet). As a final example, L14 asserts that, if the value given for 'an_8_gcps_1' is between 1 and 10 while the value for q3 is 0, then the value for the former is unjustifiably present (the corresponding ICE is thus *disinformation* rather than information), which, in this case, is dictated by the coding guidelines for the corresponding pair of questions.

15.4.2 Detailed analysis of datasets

15.4.2.1 The German Dataset

390 patients were found to have an average of 151 field names (94% of German dataset 161 variables, a field name corresponding to a value being provided for a variable), 7.2 Justified Absences (4.5%), and 4.1 Unjustified Absences (2.6%). In addition, there is an average of 1.6 Redundant Presences (1 %) per patients (Table 1).

Field Name analysis of 390 Patients

	Template Total	Mean (SD)	Pct. of Template	Min	Max
Field Name	161	151.0 (8.67)	93.80 (5.38)	87	160
Category JA: Justified Absent		7.22 (5.20)	4.48 (3.22)	0	23
Category UA: Unjustified Absent		4.14 (6.66)	2.57 (4.14)	0	53
Category RP: Redundant Present		1.64 (1.74)	1.02 (1.08)	0	15
Category UP: Unjustified Present		0.11 (0.36)	0.07 (0.34)	0	8

Out of 390 patients in German Dataset, 13 patients have significant amount of missing data or Field Name (< 84%). These 13 outliers have an average of 120 field names (74%), 14.4 Justified Absences (7.79%), and 28.5 Unjustified Absences (17.7%).

Field Name analysis of 13 Outliers

	Template Total	Mean (SD)	Pct. of Template	Min	Max
Field Name	161	119.38 (13.67)	74.15 (8.49)	87	131
Category JA: Justified Absent		14.38 (6.97)	8.93 (4.33)	2	23
Category UA: Unjustified Absent		28.46 (12.54)	17.68 (7.79)	14	53
Category RP: Redundant Present		1.08 (1.12)	0.67 (0.69)	0	3
Category UP: Unjustified Present		0.23 (0.60)	0.14 (0.37)	0	2

Field Name analysis of 377 Patients excluding 13 outliers

	Template Total	Mean (SD)	Pct. of Template	Min	Max
Field Name	161	152.1 (6.00)	94.47 (3.72)	133	161
Category JA: Justified Absent		6.97 (4.96)	4.33 (3.08)	0	21
Category UA: Unjustified Absent		3.31 (4.44)	2.05 (2.76)	0	26
Category RP: Redundant Present		1.66 (1.76)	1.03 (1.09)	0	15
Category UP: Unjustified Present		0.10 (0.55)	.063 (0.34)	0	8

An average of 737 references can be derived from an average of 151 field names, a 'reference' being a Referent Tracking compatible representation of what the value for a variable denotes for a specific patient. References are composed of 42% of Explicit /Coded Value, 54% of Implicit, 1% of Justified Absence, 0.6% of Unjustified Absence, 0.2% of Redundant Presence, and 0.01% of Unjustified Presence.

	Mean (SD)	Pct. of Reference	Min	Max
Reference	737.28 (29.94)		538	763
Category CV: Coded Value	129.17 (5.37)	17.52 (0.27)	85	133
Category IM: Implicit	576.40 (30.55)	78.14 (1.58)	358	607
Category JA: Justified Absent	7.22 (5.20)	0.99 (0.75)	0	23
Category UA: Unjustified Absent	4.14 (6.66)	0.60 (1.06)	0	53
Category RP: Redundant Present	1.64 (1.74)	0.22 (0.24)	0	15
Category UA: Unjustified Present	0.11 (0.55)	0.01 (0.07)	0	8

An average of 736 references corresponds to an average of 736 IUI for Label, i.e. IUI(L). After accounting for perspective of patient side interpretation, i.e. IUI(P), the number of IUI will be

reduced to an average of 535 IUI for Patient side particulars (IUI(P), 73% of IUI(L)). Such reduction is due to an average of 45.5 highly used IUI(P).

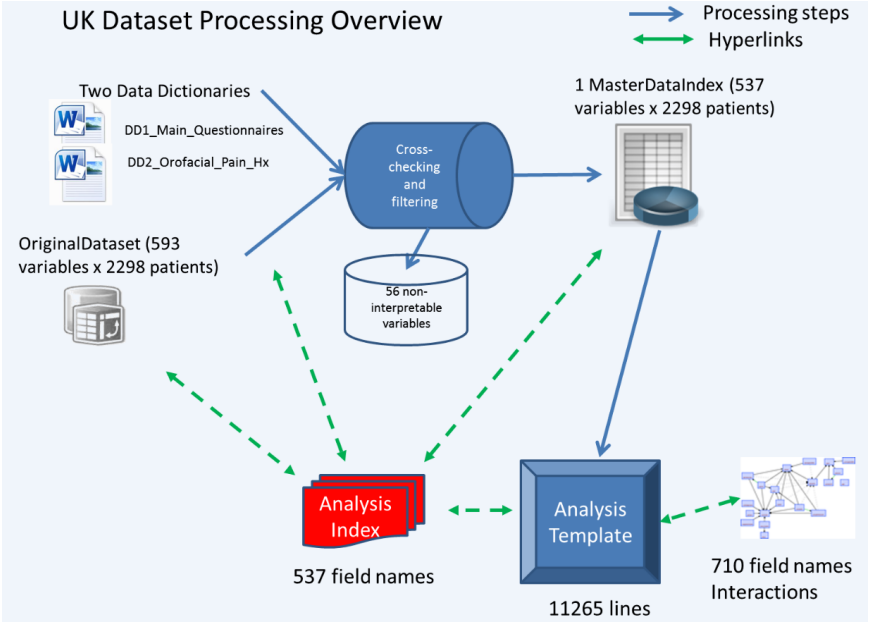
	Mean (SD)	Pct. of IUI(L)	Min	Max
IUI for Label (IUI(L))	736.37 (29.93)		537	762
IUI for Patient side (IUI(P))	535.09 (17.24)	72.71 (1.1)	408	548
Highly Used IUI(P)	45.51 (3.30)	6.18 (0.34)	24	59

15.4.2.2 Swedish dataset

For the Swedish dataset 125 subjects and 203 variables were processed. The subjects can be categorized into 3 Diagnosis types – AO case, Control, and TMD case (n=46, 38 and 41 respectively). Each diagnosis type displayed a distinct data pattern.

	AO case	Control	TMD case
Sample Size (% of total sample =125)	46 (23%)	38 (19%)	41 (20%)
Avg. Missing Data (% of total var = 204)	9.2/203 (4.9%)	44/203 (21.7%)	38/203 (19.2%)
Avg. UA (Unjustified Absent)	3.6 / 9.2 (39%)	9.6 / 44 (22%)	6.2 / 38 (16%)
Avg. JA (Justified Absent)	5.7 / 9.2 (61%)	34 / 44 (78%)	32 / 38 (84%)
Avg. Existing Data (% of total var = 204)	194/203 (95%)	159/203 (78%)	165/203 (81%)
Avg. RP(Redundant Present)	2.4 / 194 (1.3%)	3.5 / 159 (2.2%)	2 / 165 (1.2%)
Avg. UP (Unjustified Present)	1 / 194 (0.5%)	2 / 159 (1.3%)	0 (0%)

15.4.2.3 UK2 dataset



15.5 Conclusion

We have outlined a methodology for translating a clinical research dataset into a collection of Referent Tracking Tuples in such a way that both the portion of reality described by the dataset as well as the dataset itself and associated interrelations are represented in a way that mimics the structure of reality. Applying the methodology to a concrete dataset and performing some basic exploratory statistics (Table 3) revealed that there are indeed various ways in which data items can relate to what they are about (if they are about anything at all).

A set of RTTs of this sort may in the future replace the overly complicated exchange information models that are currently used in message-based paradigms [124] or in the ETL (Extract – Transform – Load) analyses and procedures common in data warehousing. To achieve the vision of maximally self-explanatory and explicit data repositories, several issues need further investigation. Although the syntax and semantics of RTTs seems powerful enough to represent what is required, a current limitation is the insufficient development of the Information Artifact Ontology. We need, above all, an adequate set of relations for the various flavors of aboutness and a better theory of Information Content Entities (ICEs), for instance concerning the various types of ICEs that exist, and how they relate to concretizations and to each other. A second limitation is that not all RTTs can easily be translated into OWL-based languages and processed by current reasoners. This is not least because of the time-dependent relationships which some RTTs involve. Whereas the former issue is a task for ontologists, the latter is to be addressed by computer scientists. Finally, building templates as described here is still very labor intensive, though we anticipate that, as experience in applying the method grows, ways will be found to automate a considerable portion of the effort involved.

16 Acknowledgements

I wish to thank the National Institute of Dental and Craniofacial Research for the financial support provided to conduct the research described in this report.

My gratitude goes also to my collaborators in this research, and in the first place to **Dr. Richard Ohrbach**, Director of the Center for Orofacial Pain Research, SMBS, University at Buffalo, whose research group focuses on a number of intersecting questions, which revolve around pain, function, diagnosis, and classification. His research activities include determinants of motor behavior and its peripheral organization, diagnostic methods, disability and the biopsychosocial model of disease, instrument development, international studies, joint mechanics, longitudinal studies and taxonomic development. Dr. Ohrbach gave the incentive for the work reported on here, provided the axis 2 data and supporting documents of the US dataset and was instrumental in providing access to major events in the domain of pain research.

I wish further to thank (listed in alphabetical order):

Dr. Vishal Aggarwal, NIHR Clinician Scientist, School of Dentistry, University of Manchester, who kindly provided the UK2 dataset. Dr. Aggarwal is an expert in disability and classification criteria for chronic orofacial pain (OFP). Using these criteria he successfully completed the first population based prospective study of OFP to investigate its etiology. He also gained broad experience in health services research particularly qualitative methods, health economics, Evidence based population health and Emergency Humanitarian Assistance.

Dr. Rafael Benoliel, originally in The Department of Oral Medicine, Faculty of Dentistry, Hebrew University Hadassah, Jerusalem 91010, Israel, and since 2013 at the Center for Orofacial Pain and Temporomandibular Disorders at Rutgers School of Dental Medicine where he also serves as the Associate Dean for Research. I am grateful for his contribution of the 'Hadassah set' as well as for specific assistance with regards to the interpretation of its variables. Dr Benoliel has published extensively on the subject of orofacial pain with over 120 papers and chapters. He co-edited the award winning textbook "Orofacial Pain and Headache' published in 2008 and now in its second edition. Dr Benoliel lectures extensively in national and international meetings. He serves on the editorial board of several leading journals and has served in many scientific committees including the Classification Committees of the International Headache Society and the Research Diagnostic Criteria for Temporomandibular Disorders.

Dr. Will Hsu who started as a PhD student in Neuroscience in UB's School for Medicine and Biomedical Sciences (SMBS), and obtained his degree in the course of the project. He was trained in the collection and analysis of physiological reactivity data from asthmatic children placed under laboratory stress protocols. He has also taken a major role in constructing, converging and maintaining physiological, psychological, and medical databases from paper charts to electronic formats. He expanded his background in Medical/Health Informatics by undertaking study in the SMBS certification program. He implemented the tools designed in this project and performed the first analyses of the materials obtained.

Dr. Mike T. John who started working as an Assistant Professor in the Department of Prosthodontics and joined the Department of Prosthodontics and Materials Sciences at the University of Leipzig, Germany as an Associate Professor in 2004. In 2007, he joined the Division of TMD and Orofacial Pain, Department of Diagnostic and Biological Sciences at the University of Minnesota. His research fields of interest include the investigation of the etiology, diagnosis and classification of temporomandibular disorders and the assessment of outcomes

of common oral treatments using the concept oral health-related quality of life. I am indebted to him for the timely delivery of the 'German Dataset' and all supporting documents.

Dr. Thomas List, Faculty of Odontology, Malmö University, Sweden whose research in recent years has been directed to the understanding of mechanisms of atypical odontalgia. He kindly provided the 'Swedish dataset' with patients with atypical odontalgia and assisted in the interpretation of its structure to facilitate the development of a classification system for Orofacial pain. He has had extensive experience in many international multi-site studies, with roles as clinical examiner, and with primary roles as both Principal Investigator and Co-Investigator

Dr. Ambra Michelotti, associate professor in Clinical Gnathology at the University of Naples Federico II, Italy. Her clinical interests are exclusively to the treatment of temporomandibular disorders and to the orthodontic practice. Dr Michelotti is an active member of, for instance, IADR (International Association of Dental Research) and of IASP (International Association of Study of Pain). She has contributed enormously to making our research more widely known in the field and became a strong advocate of the ontological approach.

Dr Donald Nixdorf, Associate Professor and Graduate Program Director at the University of Minnesota in the Division of TMD & Orofacial Pain, adjunct Assistant Professor in the Department of Neurology, Research Investigator in the HealthPartners Institute for Research and Education, and the Deputy Director of the Midwest Region within the National Dental Practice-Based Research Network (NDPBRN). I am grateful for his willingness to test the applicability of our ontological principles to improve the definition and characterization of orofacial pain disease entities.

Dr. Eric Schiffman, associate professor and Director of the Division of TMD and Orofacial Pain at the School of Dentistry at the University of Minnesota, for delivering the axis I part of the US dataset.

Dr. Joanna M. Zakrzewska, Facial pain unit, Division of Diagnostic, Surgical and Medical Sciences, Eastman Dental Hospital, UCLH NHS Foundation Trust, London, UK, for providing the UK1 dataset. Her research on orofacial pain and particularly trigeminal neuralgia is clinically based and she works not only with multidisciplinary teams at UCLH, London but also with teams nationally and internationally. Her key areas of research are clinical trials in trigeminal neuralgia, systematic reviews and guidelines on management of trigeminal neuralgia and patient involvement in trigeminal neuralgia studies.

17 References

1. Maojo, V., et al., *Biomedical Ontologies: Toward Scientific Debate.* Methods of Information in Medicine, 2011. **50**(3): p. 203-216.
2. Brochhausen M, et al., *Discussion of "Biomedical Ontologies: Toward Scientific Debate".* Methods of Information in Medicine, 2011. **50**(3): p. 217-236.
3. Nixdorf D, et al., *Classifying orofacial pains: a new proposal of taxonomy based on ontology.* Journal of Oral Rehabilitation, 2012. **39**(3): p. 161-169.
4. Ioannidis, J. *Why most published research findings are false*. PLoS Med, 2005. **2**, e124.
5. Young, N., J. Ioannidis, and O. Al-Ubaydli *Why Current Publication Practices May Distort Science*. PLoS Med, 2008. **5**, e201 DOI: doi:10.1371/journal.pmed.0050201.
6. Ceusters, W., Smith, B., *Tracking Referents in Electronic Health Records*, in *Connecting Medical Informatics and Bio-Informatics: Proceedings of MIE 2005 The XIXth International Congress of the European Federation for Medical Informatics*, G.A. Engelbrecht R., Lovis, C., Editor. 2005, IOS Press: Amsterdam. p. 71-76.
7. Ceusters, W., Steurs, F., Zanstra, P., Van Der Haring, E., Rogers, J., *From a Time Standard for Medical Informatics to a Controlled Language for Health.* International Journal of Medical Informatics, 1998. **48**(1-3): p. 85-101.
8. Ceusters, W., Smith, B., *Strategies for referent tracking in electronic health records.* Journal Biomedical Informatics, 2006. **39**(3): p. 362-378.
9. Ceusters, W., P. Elkin, and B. Smith, *Referent Tracking: The Problem of Negative Findings*, in *Studies in Health Technology and Informatics. Ubiquity: Technologies for Better Health in Aging Societies - Proceedings of MIE2006*, A. Hasman, et al., Editors. 2006, IOS Press: Amsterdam. p. 741-746.
10. Ceusters, W., P. Elkin, and B. Smith, *Negative Findings in Electronic Health Records and Biomedical Ontologies: A Realist Approach.* International Journal of Medical Informatics, 2007. **76**: p. 326-333.
11. Ceusters, W., et al., *Managed convergence towards high quality electronic healthcare records in Europe : the PROREC initiative*. 1996, Medical Records Institute: Newton, MA. p. 127-136.
12. Buekens, F., W. Ceusters, and G.D. Moor, *The explanatory role of events in causal and temporal reasoning in medicine.* Methods of Information in Medicine, 1993. **32**: p. 274-278.
13. Ceusters, W., *Formal terminology management for language-based knowledge systems: resistance is futile*, in *Trends in Special Language and Language Technology*, R. Temmerman and M. Lutjeharms, Editors. 2001, Uitgeverij De Boeck: Antwerpen. p. 135-153.
14. Smith, B., Ceusters, W.,. *An Ontology-Based Methodology for the Migration of Biomedical Terminologies to Electronic Health Records*. in *AMIA 2005*. 2005. Washington, DC.
15. Ceusters, W. and B. Smith, *A Realism-Based Approach to the Evolution of Biomedical Ontologies*, in *Biomedical and Health Informatics: Proceedings of the 2006 AMIA Annual Symposium*. 2006, American Medical Informatics Association: Washington DC. p. 121-125.
16. Rosse, C., Mejino, J. , *A reference ontology for biomedical informatics: the Foundational Model of Anatomy. .* J Biomed Inform. , 2003. **36**(6): p. 478-500.
17. Rosse, C. and M.J. Jr, *The Foundational Model of Anatomy Ontology*, in *Anatomy Ontologies for Bioinformatics: Principles and Practice*, A. Burger, D. Davidson, and R. Baldock, Editors. 2007, Springer: London. p. 59-117.

18. Rosse, C. and J.L.V. Mejino, *A reference ontology for bioinformatics: The Foundational Model of Anatomy.* Journal of Biomedical Informatics, 2003. **36**: p. 478-500.
19. Ceusters, W., Smith, B., Kumar A., Dhaen C. *Ontology-Based Error Detection in SNOMED-CT®*. in *Proceedings of Medinfo*. 2004.
20. Elhanan, G., Y. Perl, and J. Geller, *A Survey of Direct Users and Uses of SNOMED CT: 2010 Status.*, in *AMIA Annu Symp Proc*. 2010 p. 207-11.
21. Elkin, P., et al., *Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists.* Mayo Clin Proc, 2006. **81**(6): p. 741-748.
22. Rector, A., *Terminologies, Ontologies, & SNOMED. What are they for? What would Quality Assurance mean?*, in *First European Conference on SNOMED CT*. 2006: Copenhagen.
23. Ceusters, W., B. Smith, and L. Goldberg, *A terminological and ontological analysis of the NCI Thesaurus.* Methods of Information in Medicine, 2005. **44**: p. 498-507.
24. Coronado, S.d., et al., *The NCI Thesaurus quality assurance life cycle.* Journal of Biomedical Informatics, 2009. **42**(3): p. 530-539.
25. de Coronado, S., Haber, M., Sioutos, N., Tuttle, M., Wright, L. *NCI Thesaurus: Using Science-Based Terminology to Integrate Cancer Research Results*. in *Medinfo*. 2004. IOS Press.
26. Kumar, A. and B. Smith, *Oncology ontology in the NCI Thesaurus*, in *Artificial Intelligence in Medicine Europe (Lecture Notes in Computer Science 3581)*. 2005. p. 213-220.
27. Schulz, S., et al., *The Pitfalls of Thesaurus Ontologization - the Case of the NCI Thesaurus*, in *AMIA Annual Symposium Proceedings*. 2010, AMIA: Washington D.C. p. 727-731.
28. Sioutos, N., et al., *NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information.* Journal of Biomedical Informatics 40, 2007. **40**(1): p. 30-43.
29. Ceusters, W., *Applying Evolutionary Terminology Auditing to the Gene Ontology.* Journal of Biomedical Informatics; Special Issue of the Journal of Biomedical Informatics on Auditing of Terminologies, 2009. **42**(3): p. 518-529.
30. Ohrbach, R., et al., *Recommendations from the International Consensus Workshop: Convergence on an Orofacial Pain Taxonomy.* Journal of Oral Rehabilitation, 2010.
31. Scheuermann, R.H., W. Ceusters, and B. Smith, *Toward an Ontological Treatment of Disease and Diagnosis*, in *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics, San Francisco, California, March 15-17, 2009*. 2009, American Medical Informatics Association. p. 116-120.
32. National Cancer Institute. *NCIthesaurus: Finding (Code C3367)*. 2009; Available from: http://ncit.nci.nih.gov/ncitbrowser/ConceptReport.jsp?dictionary=NCI%20Thesaurus&code=C3367.
33. Loscalzo, J., I. Kohane, and A.-L. Barabasi, *Human disease classification in the postgenomic era: A complex systems approach to human pathobiology.* Mol Syst Biol, 2007. **3**(124-134).
34. Butte, A. and I. Kohane, *Creation and implications of a phenome-genome network.* Nat Biotechnol, 2006. **24**(1): p. 55-62.
35. Schulz, S. and I. Johansson, *Continua in biological systems.* The Monist, 2007. **90**(4): p. 499-522.
36. Goldfain, A. *Ontology for General Medical Science (OGMS)*. 2009 [cited 2009 September 13]; Available from: http://www.acsu.buffalo.edu/~ag33/ogms.html.
37. Smith, B., et al., *Towards an Ontology of Pain and of Pain-Related Phenomena.* 2009: p. (submitted to MEDINFO 2009).

38. *Stohler, C.S.*, *The End of an Era: Orofacial Pain Enters the Genomic Age. Implications and Opportunities for Research and the Care of Patients*, in *The Puzzle of Orofacial Pain. Integrating Research into Clinical Management. Pain Headache.*, Türp JC, Sommer C, and Hugger A, Editors. 2007, Karger: Basel. p. 236–247.
39. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nature Biotechnology, 2007. **25**: p. 1251-1255.
40. Smith, B., et al., *Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain*, in *KR-MED 2006, Biomedical Ontology in Action.* 2006: Baltimore MD, USA
41. Ceusters, W. and B. Smith *Referent Tracking and its Applications* CEUR Workshop Proceedings, 2007. **249**.
42. International RDC-TMD Consortium Network. *International RDC/TMD Consortium: A Designated Network of the International Association for Dental Research*. 2009 [cited 2009 Sept 10]; Available from: http://www.rdc-tmdinternational.org/.
43. International Association for the Study of Pain. *Special Interest Group: Orofacial Pain*. 2009 [cited 2009 Sept 10]; Available from: http://www.iasp-pain.org/AM/Template.cfm?Section=SIGS&Template=/CM/HTMLDisplay.cfm&ContentID=1594.
44. Ohrbach, R., *Disability Assessment and Management in Rehabilitation of the Masticatory System*, in *Colloquium on Oral Rehabilitation 2009*. 2009: Siena, Italia.
45. World Health Organization, *International Classification of Impairments, Disabilities, and Handicaps.* 1980, Geneva:: World Health Organization.
46. World Health Organization, *International Classification of Functioning, Disability and Health (ICF)*. 2002, Geneva: World Health Organization.
47. Institute of Medicine's Committee on Pain Disability and Chronic Illness Behavior, *Pain and disability : clinical, behavioral, and public policy perspectives*, ed. Osterweis M, Kleinman A, and Mechanic D. 1987, Washington, D.C: National Academy Press.
48. Merskey H, et al., *Pain terms: A list with definitions and notes on usage; recommended by the IASP Subcommittee on Taxonomy.* Pain, 1979. **6**: p. 249-252.
49. International Association for the Study of Pain. *IASP Pain Taxonomy*. 2012; Available from: http://www.iasp-pain.org/Content/NavigationMenu/GeneralResourceLinks/PainDefinitions/default.htm.
50. Melzack R, *From the gate to the neuromatrix.* Pain, 1999. **Supplement 6**: p. S121-S126.
51. Dworkin, S., *Somatization, Distress and Chronic Pain* Quality of Life Research 1994. **3, Supplement: Chronic Pain** (1): p. S77-S83.
52. *Schatman, M.E.*, *The Challenge of The Dramatically Disturbed Chronic Pain Patient. The Pain Practitioner*, 2003. **13**(1): p. 5-7.
53. National Institutes of Health, *Technology Assessment Conference Statement: management of temporomandibular disorders.* Oral Surg Oral Med Oral Pathol Oral Radiol Endod, 1997. **83**: p. 177-183.
54. Suvinen, T.I., et al., *Review of aetiological concepts of temporomandibular pain disorders: towards a biopsychosocial model for integration of physical disorder factors with psychological and psychosocial illness impact factors.* European Journal of Pain, 2005. **9**: p. 613-633.
55. Dworkin, S.F. and L. LeResche, *Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications.* Journal of Craniomandibular Disorders, 1992. **6**(4): p. 301-355.
56. Manfredini D and Guarda-Nardini L, *Agreement between Research Diagnostic Criteria for Temporomandibular Disorders and Magentic Resonance Diagnoses of*

Temporomandibular disc displacement in a patient population. International Journal of Oral and Maxillofacial Surgery, 2008. **37**(7): p. 612-616.
57. Walker JG, Jackson HJ, and Littlejohn GO, *Models of adjustment to chronic illness: using the example of rheumatoid arthritis.* Clin Psychol Rev, 2004. **24**: p. 461-488.
58. John, M.T., et al., *Oral health-related quality of life in patients with temporomandibular disorders.* Journal of Orofacial Pain, 2007. **21**(1): p. 46-54.
59. Schiffman, E., et al., *Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) for Clinical and Research Applications: Recommendations of the International RDC/TMD Consortium Network and Orofacial Pain Special Interest Group.* Journal of Oral and Facial Pain and Headache, 2014. **28**(1): p. 6-27.
60. Slade GD and Spencer AJ, *Development and evaluation of the Oral Health Impact Profile.* Community Dental Health, 1994. **11**(1): p. 3-11.
61. Mancl, L., C. Whitney, and X. Zhu, *A SAS computer program to evaluate the research diagnostic criteria for classification of temporomandibular disorders*, in *Technical Report Series*. 1999, University of Washington. p. 44.
62. Benoliel, R., E. Eliav, and Y. Sharav, *Self Reports of Pain-Related Awakenings in Persistent Orofacial Pain Patients.* Journal of Orofacial Pain, 2009. **23**(4): p. 1-9.
63. Ahmad, M., et al., *Research diagnostic criteria for temporomandibular disorders (RDC/TMD): development of image analysis criteria and examiner reliability for image analysis.* Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, 2009. **107**(6): p. 844-860.
64. Ceusters, W., K.A. Spackman, and B. Smith. *Would SNOMED CT benefit from Realism-Based Ontology Evolution?* in *American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy*. 2007. Chicago IL: American Medical Informatics Association.
65. Cimino, J.J., *Desiderata for controlled medical vocabularies in the twenty-first century.* Methods of Information in Medicine, 1998. **37**(4-5): p. 394-403.
66. Boeker, M., et al., *Unintended consequences of existential quantifications in biomedical ontologies.* BMC Bioinformatics, 2011. **12**: p. 456.
67. Guarino, N. and P. Giaretta, *Ontologies and Knowledge Bases: Towards a Terminological Clarification*, in *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, N. Mars, Editor. 1995, IOS Press: Amsterdam. p. 25-32.
68. Bodenreider, O., *Medical Ontology Research: A Report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications*. 2001.
69. Smith, B. and W. Ceusters, *HL7 RIM: An Incoherent Standard*, in *Studies in Health Technology and Informatics. Ubiquity: Technologies for Better Health in Aging Societies - Proceedings of MIE2006*, A. Hasman, et al., Editors. 2006, IOS Press: Amsterdam. p. 133-138.
70. Smith, B. and W. Ceusters, *An Ontology-Based Methodology for the Migration of Biomedical Terminologies to Electronic Health Records*, in *AMIA 2005*. 2005: Washington DC. p. 669-673.
71. Bechhofer, S., et al. *OWL Web Ontology Language Reference*. 2004 [cited 2008 December 10]; Available from: http://www.w3.org/TR/owl-ref/.
72. Manzano, M., *Model Theory*. 1999, Oxford: Oxford University Press.
73. Baud, R., et al., *Reconciliation of Ontology and Terminology to cope with Linguistics*, in *Proceedings of MEDINFO 2007, Brisbane, Australia, August 2007*, K. Kuhn, J. Warren, and T. Leong, Editors. 2007, Ios Press: Amsterdam. p. 796-801.

74. Smith, B. and W. Ceusters, *Towards Industrial-Strength Philosophy; How Analytical Ontology Can Help Medical Informatics.* Interdisciplinary Science Reviews, 2003. **28**(2): p. 106-111.
75. Ceusters, W. and B. Smith, *Strategies for Referent Tracking in Electronic Health Records.* Journal of Biomedical Informatics, 2006. **39**(3): p. 362-378.
76. Manzoor, S., W. Ceusters, and R. Rudnicki, *Implementation of a Referent Tracking System.* International Journal of Healthcare Information Systems and Informatics, 2007. **2**(4): p. 41-58.
77. Rudnicki, R., et al., *What Particulars are Referred to in EHR Data? A Case Study in Integrating Referent Tracking into an Electronic Health Record Application*, in *American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy*, Teich JM, Suermondt J, and H. C, Editors. 2007: Chicago, IL. p. 630-634.
78. Ceusters, W. and B. Smith, *Referent Tracking for Treatment Optimisation in Schizophrenic Patients.* Journal of Web Semantics - Special issue on semantic web for the life sciences, 2006. **4**(3): p. 229-36.
79. Smith, B. and W. Ceusters, *Ontological realism: A methodology for coordinated evolution of scientific ontologies.* Appl Ontol, 2010. **5**(3-4): p. 139-188.
80. Ceusters, W. and B. Smith, *Foundations for a realist ontology of mental disease.* Journal of Biomedical Semantics, 2010. **1**(10): p. 1-23.
81. Hastings, J., et al., *Representing Mental Functioning: Ontologies for Mental Health and Disease*, in *Towards an Ontology of Mental Functioning (ICBO Workshop), Proceeedings of the Third International Conference on Biomedical Ontology*. 2012.
82. Smith, B. and W. Ceusters, *Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies.* Applied Ontology, 2010. **5**(3-4): p. 139-188.
83. Schulz, S., M. Brochhausen, and R. Hoehndorf, *Higgs Bosons, Mars Missions, and Unicorn Delusions: How to Deal with Terms of Dubious Reference in Scientific Ontologies*, in *Proceedings of the International Conference on Biomedical Ontologies.*, B. Smith, Editor. 2011: Buffalo NY,. p. 183-189.
84. Ceusters, W., *An information artifact ontology perspective on data collections and associated representational artifacts.* Stud Health Technol Inform, 2012. **180**: p. 68-72.
85. Klaus, K., et al., *The Distinction Between "Medically Unexplained" and "Medically Explained" in the Context of Somatoform Disorders.* Int J Behav Med, 2012.
86. Huang, H. and R.M. McCarron, *Medically unexplained physical symptoms: Evidence-based interventions.* Current Psychiatry, 2011. **10**(7): p. 17.
87. Carruthers, B.M., et al., *Myalgic encephalomyelitis/chronic fatigue syndrome.* Journal of chronic fatigue syndrome, 2003. **11**(1): p. 7-115.
88. Greco, M., *The classification and nomenclature of 'medically unexplained symptoms': conflict, performativity and critique.* Soc Sci Med, 2012: p. 1-31.
89. Doing-Harris, K., et al., *Applying ontological realism to medically unexplained syndromes.* Stud Health Technol Inform, 2013. **192**: p. 97-101.
90. Bodenreider, O., B. Smith, and A. Burgun, *The ontology-epistemology divide: A case study in medical terminology*, in *Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004)*, A.C. Varzi and L. Vieu, Editors. 2004, IOS Press: Amsterdam. p. 185-195.
91. Noy, N.F., et al., *BioPortal: ontologies and integrated data resources at the click of a mouse.* Nucleic Acids Research, 2009 **1**: p. 37.
92. Jonquet, C., N.H. Shah, and M.A. Musen, *The open biomedical annotator.* Summit on Translat Bioinforma, 2009. **2009**: p. 56-60.

93. Ceusters, W. *Supplementary data to Pain Assessment Terminology in the NCBO BioPortal: Evaluation and Recommendations.* 2014.
94. Whetzel, P.L. and N. Team, *NCBO Technology: Powering semantically aware applications.* J Biomed Semantics, 2013. **4 Suppl 1**: p. S8.
95. Cimino, J.J., *In Defense of the desiderata.* Journal of Biomedical Informatics, 2006. **39**(3): p. 299-306.
96. Schulz, S. and L. Jansen, *Formal ontologies in biomedical knowledge representation.* Yearb Med Inform, 2013. **8**(1): p. 132-46.
97. Smith B, et al., *Towards an Ontology of Pain*, in *Proceedings of the Conference on Logic and Ontology*, M. Okada, Editor. 2011, Keio University Press: Tokyo. p. 23-32.
98. Ruttenberg, A. and R. Ferguson. *[bioontology-support] Something amiss in translation of WHOART*. 2011 [cited 2014 April 15]; Available from: https://mailman.stanford.edu/pipermail/bioontology-support/2011-April/003124.html.
99. Schulz, S., et al., *From concept representations to ontologies: A paradigm shift in health informatics?* Healthcare Informatics Research, 2013. **19**(4): p. 235-242.
100. He, Z., et al., *A family-based framework for supporting quality assurance of biomedical ontologies in BioPortal.* AMIA Annu Symp Proc, 2013. **2013**: p. 581-90.
101. Welch, W.H., *Papers and addresses*. Vol. 3. 1822, Baltimore: The John Hopkins Press.
102. IASP Subcommittee on Taxonomy, *Pain terms: a list with definitions and notes on usage.* Pain, 1979. **6**(3): p. 249-252.
103. Merskey H and Bogduk N, *Classifications of chronic pain: Description of chronic pain syndromes and definition of pain terms. Report by the International Association for the Study of Pain Task Force on Taxonomy*. 1994, Seattle: IASP Press.
104. Smith, B., *Introduction to the Logic of Definitions*, in *International Workshop on Definitions in Ontologies, organized in conjunction with the Fourth International Conference on Biomedical Ontology (ICBO)*. 2013, CEUR: Montreal. p. 1-2.
105. Turk, D.C. and R. Melzack, *Handbook of pain assessment*. 3rd ed. 2011, New York: Guilford Press. xvii, 542 p.
106. Ceusters, W., *Pain assessment terminology in the NCBO BioPortal: evaluation and recommendations*, in *Proceedings of the International Conference on Biomedical Ontology 2014*. 2014: Houston, TX. p. (accepted).
107. Smith, B. *Basic Formal Ontology 2.0*. 2012 [cited 2012 January 24]; Available from: http://ontology.buffalo.edu/bfo/Reference/.
108. Smith, B. and W. Ceusters, *Ontological Realism: A Methodology for Coordinated Evolution of Scientific Ontologies.* Applied Ontology, 2010. **5**: p. 139-188.
109. Smith, B., *On classifying Material Entities in Basic Formal Ontology*, in *Interdisciplinary Ontology. Proceedings of the Third Interdisciplinary Ontology Meeting*. 2012. p. 1-13.
110. Smith, B., et al., *Basic Formal Ontology 2.0: DRAFT SPECIFICATION AND USER'S GUIDE*. 2012.
111. Haas, L., *Beauty and the Beast: The Theory and Practice of Information Integration*, in *Lecture Notes in Computer Science*, T. Schwentick and D. Suciu, Editors. 2007, Springer-Verlag Berlin, Heidelberg. p. 28-43.
112. Marenco, L., R. Wang, and P. Nadkarni, *Automated Database Mediation Using Ontological Metadata Mappings.* J Am Med Inform Assoc, 2009. **16**(5): p. 723-737.
113. Sim, I., et al., *Ontology-Based Federated Data Access to Human Studies Information*, in *AMIA Annu Symp Proc 2012*, . Editor. 2012: Chicago IL. p. 856-865.
114. Baumbach, J., et al., *CoryneRegNet: an ontology-based data warehouse of corynebacterial transcription factors and regulatory networks.* BMC Genomics, 2006. **7**: p. 24.

115. Rodriguez-Muro, M. and D. Calvanese, *Dependencies: Making Ontology Based Data Access Work In Practice.* , in *Proc. of the 5th Alberto Mendelzon Int. Workshop on Foundations of Data Management (AMW 2011).* . 2011.
116. Kohler, J., S. Philippi, and M. Lange, *SEMEDA: ontology based semantic integration of biological databases.* Bioinformatics, 2003. **19**(18): p. 2420-7.
117. Poggi, A., et al., *Linking data to ontologies*, in *Journal on data semantics X*, S. Spaccapietra, Editor. 2008, Springer-Verlag: Heidelberg. p. 133-173.
118. Ceusters, W. and S. Manzoor, *How to track Absolutely Everything?*, in *Ontologies and Semantic Technologies for the Intelligence Community. Frontiers in Artificial Intelligence and Applications.*, L. Obrst, T. Janssen, and W. Ceusters, Editors. 2010, IOS Press: Amsterdam. p. 13-36.
119. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biology, 2005. **6**(5): p. R46.
120. European Committee for Standardization, *EN 12388:2005. Health informatics - Time standards for healthcare specific problems.* 2005.
121. Russell, B., *On Denoting.* Mind, 1905. **56**: p. 479-93.
122. Hogan, W.R., et al., *Representing Local Identifiers in a Referent-Tracking System*, in *Proceedings of the International Conference on Biomedical Ontology*, B. Smith, Editor. 2011: Buffalo NY. p. 252-254.
123. Ceusters, W. and B. Smith, *A Unified Framework for Biomedical Terminologies and Ontologies*, in *Proceedings of the 13th World Congress on Medical and Health Informatics (Medinfo 2010), Cape Town, South Africa, 12-15 September 2010*, C. Safran, H. Marin, and S. Reti, Editors. 2010, IOS Press: Amsterdam. p. 1050-1054.
124. Smith, B., Vizenor L, and W. Ceusters, *Human Action in the Healthcare Domain: A Critical Analysis of HL7's Reference Information Model.* , in *Johanssonian Investigations: Essays in Honour of Ingvar Johansson on His Seventieth Birthday*, Svennerlind C, Almäng J, and Ingthorsson R, Editors. 2013, Ontos Verlag: Frankfurt. p. 554-573.

18 Appendix

Ontology Tools and Ontology DBs

- For Terminology Alignment and Application Ontology Development

User's Guide

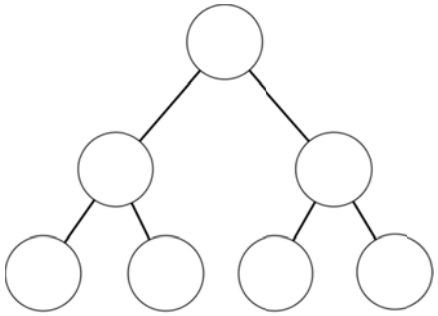


Table of Contents

A. System Overview

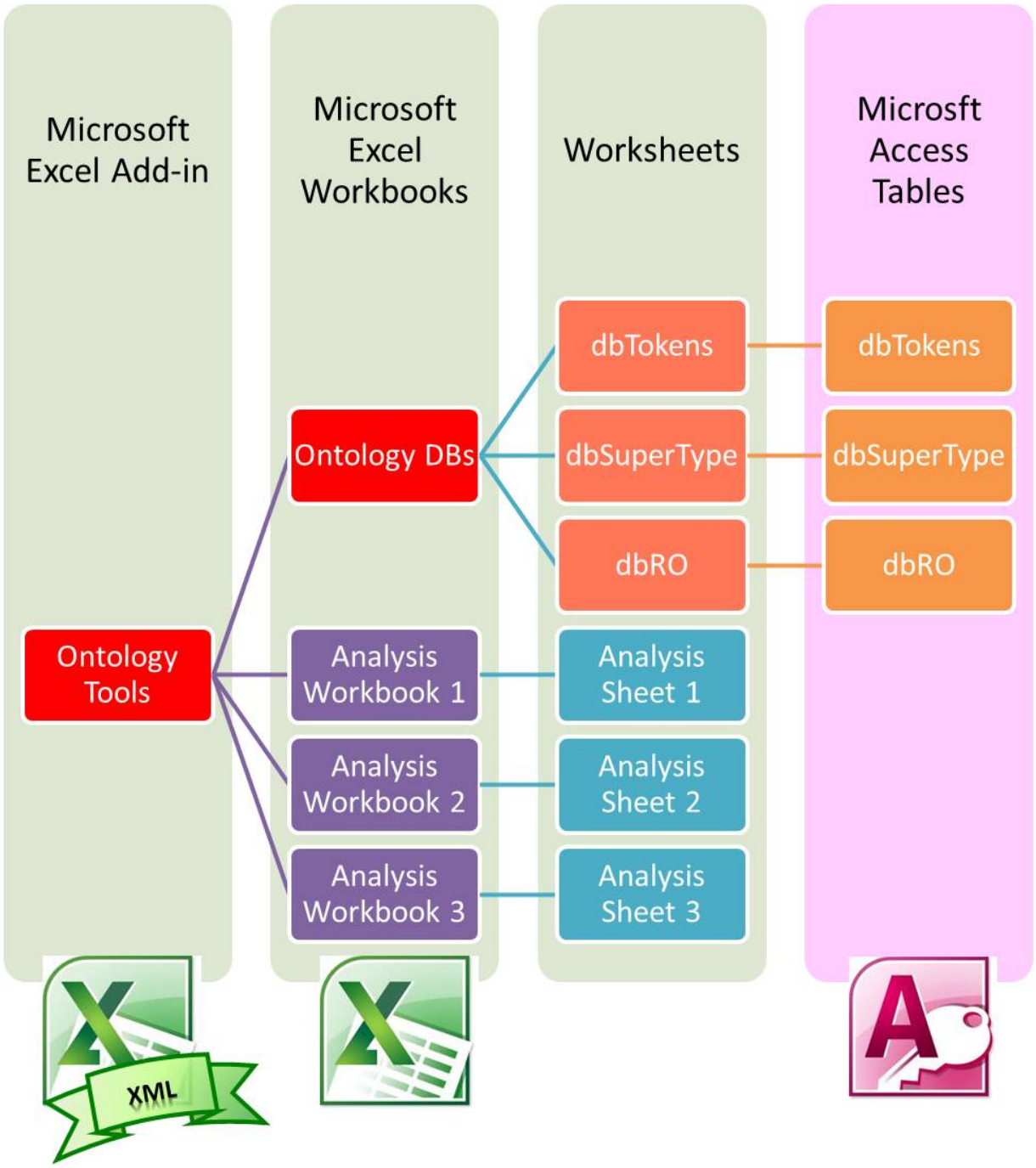


Figure 1. System Overview

Introduction

Ontology Tools and Ontology DBs are designed to facilitate the process of building application ontology. The following is the flowchart of building application ontology and how Ontology Tools facilitate the process:

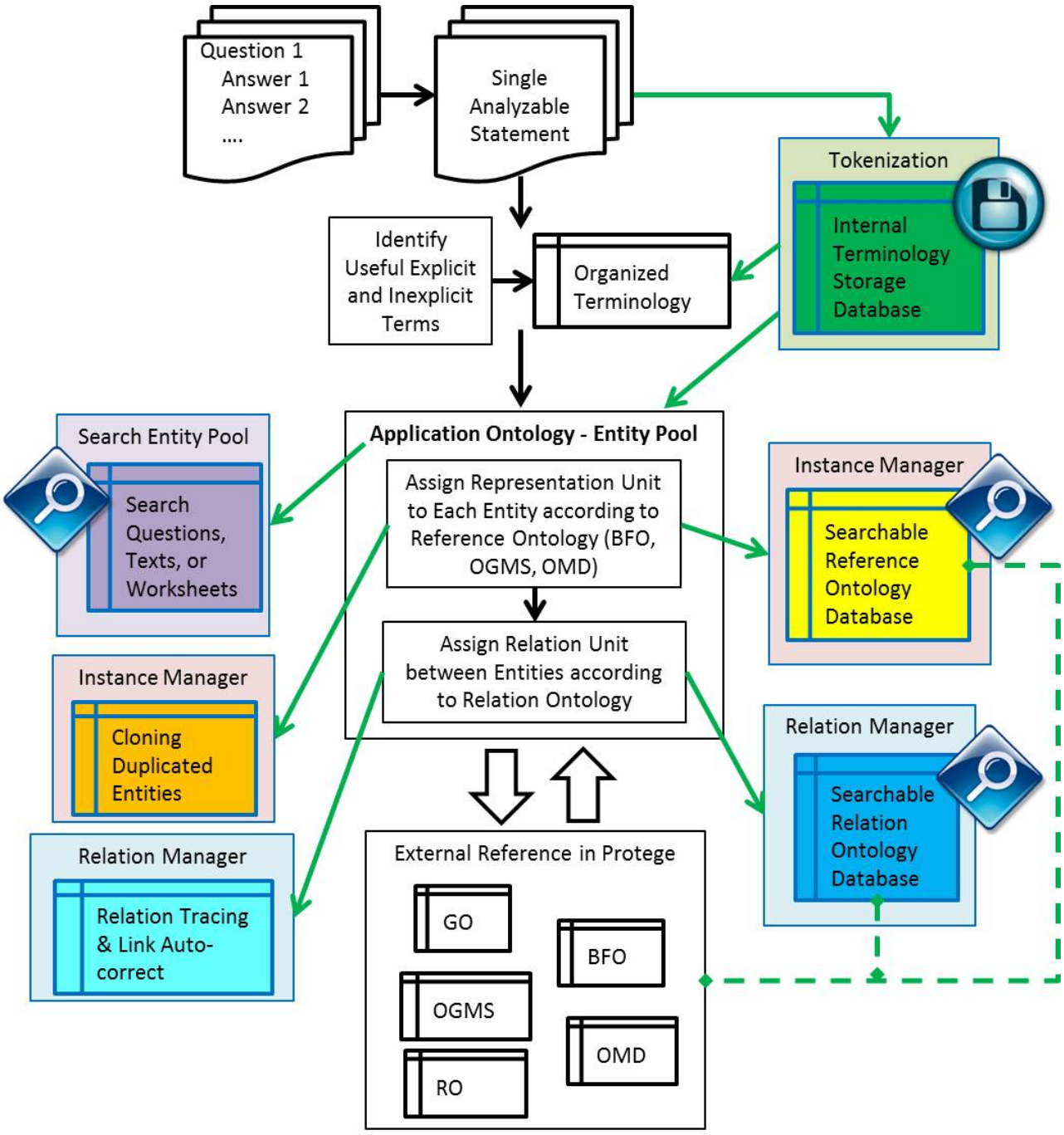


Figure 2. Building Application Ontology

All the steps in the building process can be completed in Microsoft Excel. There will be 3 components or files involved in the process, your analysis files, OntologyDB and Ontology Tools.

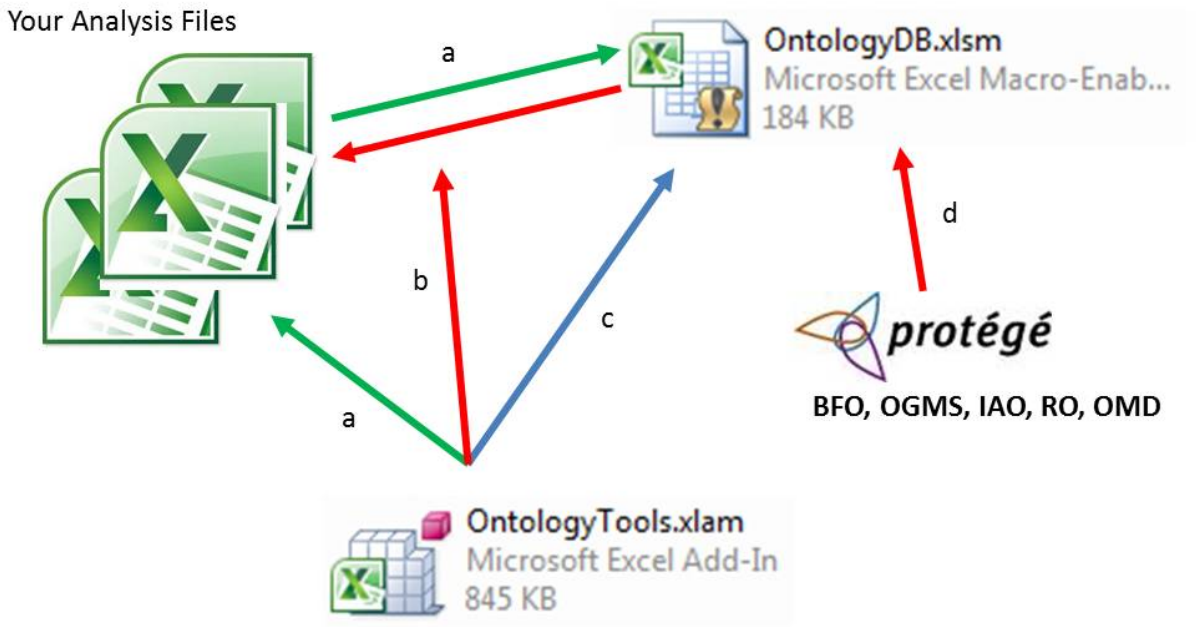


Figure 3. File Interaction Overview

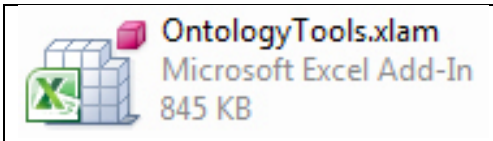
a. Facilitate breaking down statement into useful entities or terms. Meaningful terms or texts can be stored in OntologyDB.
b. Navigate and link entities and relations to established categories in Reference Ontologies.
c. Maintain and update libraries of Reference Ontologies.
d. Import Reference Ontologies into organized and searchable libraries.

Analysis files contain original texts imported from assessment instruments, such as questions and response scales. Each set of a question and a response scale will be manually converted into a single analyzable statement using realism ontology principles, such as constructing descriptive from a third person and layperson perspective, removing ambiguous pronouns and abbreviations, and restoring absent context due to purposeful omission in semantics.

OntologyDB is a collection of Reference Ontologies Libraries imported from Protégé/OWL files (Figure 3d). Reference Ontologies are organized in a searchable format and a linkable fashion from your analysis files.

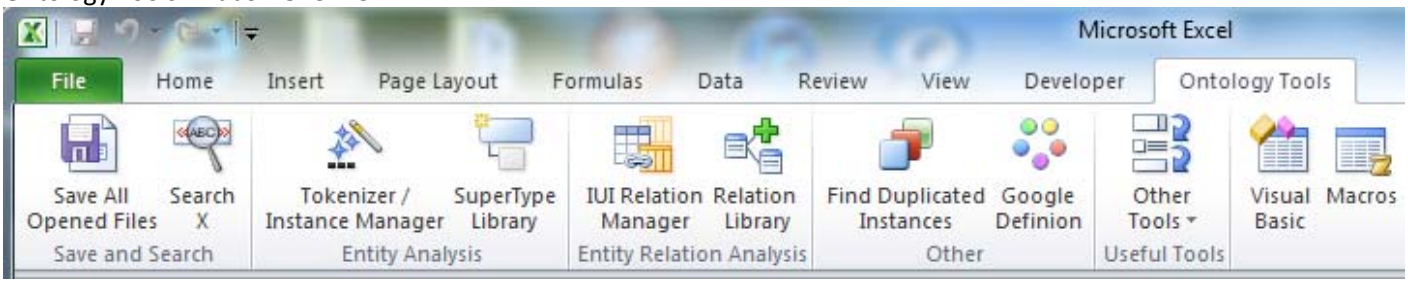
Ontology Tools is a Microsoft Excel Add-In and programmed to facilitate the process of application ontology building (Figure 3a,b,c). Please refer to next section for how each tool works.

System Basics



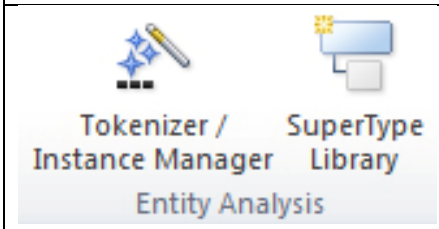
Ontology Tools is a **Microsoft Excel Add-in (.xlam)** and programmed in language of **Visual Basic Application (VBA)**. It is embedded as a form of **Microsoft Excel Customized Ribbon Menu** and programmed in **XML**. It controls multiple ontological analysis workbooks, manages connections between each analysis workbook and Ontology databases, and facilitates the processes of terminology alignment and application ontology development

Ontology Tools: Ribbon Overview



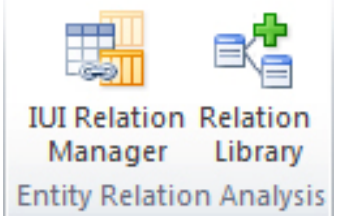
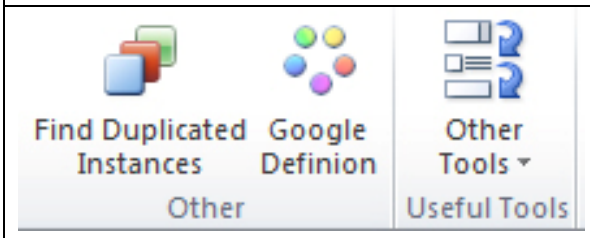
Save All Opened Files: This tool will call a Macro that can rotate through each opened Excel file and save the file bypassing Excel warning signs.

Search X: This tool will call out a UserForm/VBA Module that can 1) open, index and navigate Excel files and all the worksheets in each file, 2) index all numbered statement sections and allow to hide/show each section, and 3) search a keyword in a column, and allow to hide/show each keyword-containing section (Referred to Module Detail).



Tokenizer / Instance Manager: This tool will call out a UserForm/VBA Module with 2 major functions: Tokenizer and Instance Manager. **Tokenizer** can 1) execute tokenization process of an analyzable statement, 2) collect useful/exclusion terms into databases in OntologyDB, and 3) insert tokenized terms or tokens into a specified section under the statement. **Instance Manager** can 1) allow user to modify basic information of an entity or instance (name/supertype/description), 2) pseudo-formalize the instance description with local pool of instance IUI codes (intra-statement), 3) determine uniqueness of an instance by searching across statement sections, and 4) clone the first appearance of an instance if current instance is not unique (Referred to Module Detail).

SuperType Library: This tool will call out a UserForm/VBA Module that can 1) navigate through the SuperType Library generated from existing Reference Ontologies, 2) Add or update a SuperType category, and insert an direct formula link of a supertype to an instance currently reviewed by the **Instance Manager** (Referred to Module Detail).

	**IUI Relation Manager**: This tool will call out a UserForm/VBA Module that can 1) build a relation with 3 or 4 components (current instance (relation donor), another instance (relation receiver), temporal instance associated with relation occurring timeframe (if necessary), and relationship classification according Relation Ontology) following Relation Ontology principles, 2) translate IUI codes to instance names by tracing to active location of the instance, 3) determine uniqueness of an relation by searching across statement sections, 4) clone the first appearance of a relation if current relation is not unique, 5) trace and auto-correct the formula link of relation receiver/donor/temporal instances referencing to local pool of instances or the 1st appearance of an instance (Referred to Module Detail). **Relation Library**: This tool will call out a UserForm/VBA Module that can 1) navigate through the Relation Library generated from existing Reference Ontologies, 2) Add or update a relationship type category, and insert an direct formula link of a relationship type to an relation currently reviewed by the **IUI Relation Manager** (Referred to Module Detail).
	**Find Duplicated Instances**: This tool will call out a UserForm/VBA Module that can 1) determine uniqueness of an instance by searching across statement sections. This is view-only function isolated from instance manger (Referred to Module Detail). **Google Definition**: This tool will call out a UserForm/VBA Module that can 1) extract definitions of a single text available on Google JSTOR server. This tool requires active internet connection. **Other Tools**: This is a collection of useful Macros that will eventually be incorporated into other tools or retired.

B. Analysis File Organization

Overview

Ontology Tools currently are only functional with analysis workbook containing an analysis worksheet named “Analysis”. In this “Analysis” worksheet, a specific format is required for maximal performance of Ontology Tools as shown in figure 4. **An analysis worksheet template is available in OntologyDB file.**

Column Content (Vertical Units) The column heads define the content of each column similar to common database structure (see Table 1). Ontology Tools used the column heads to search specific contents. Columns A-C (1-3) contain important markers (A/1), question or statement numbers (B/2), and full text of a question or statement (C/3)). Columns B-Q (4-17) contain content related to entities extracted from the statements. Columns R-W (18-23) contain content related to entity relations or interaction between entities within or acrros statement sections.

Row Content (Horizontal Units) The 1st column contain the markers that define the boundaries of each statement (x>, <x) and the boundaries of entities or entity relations (h>, <h). These markers are used by Ontology Tools to 1) index statement locations, 2) determine or trace entity origin, and 3) hide/show the entity sections.

Figures 4. Analysis Worksheet Structure

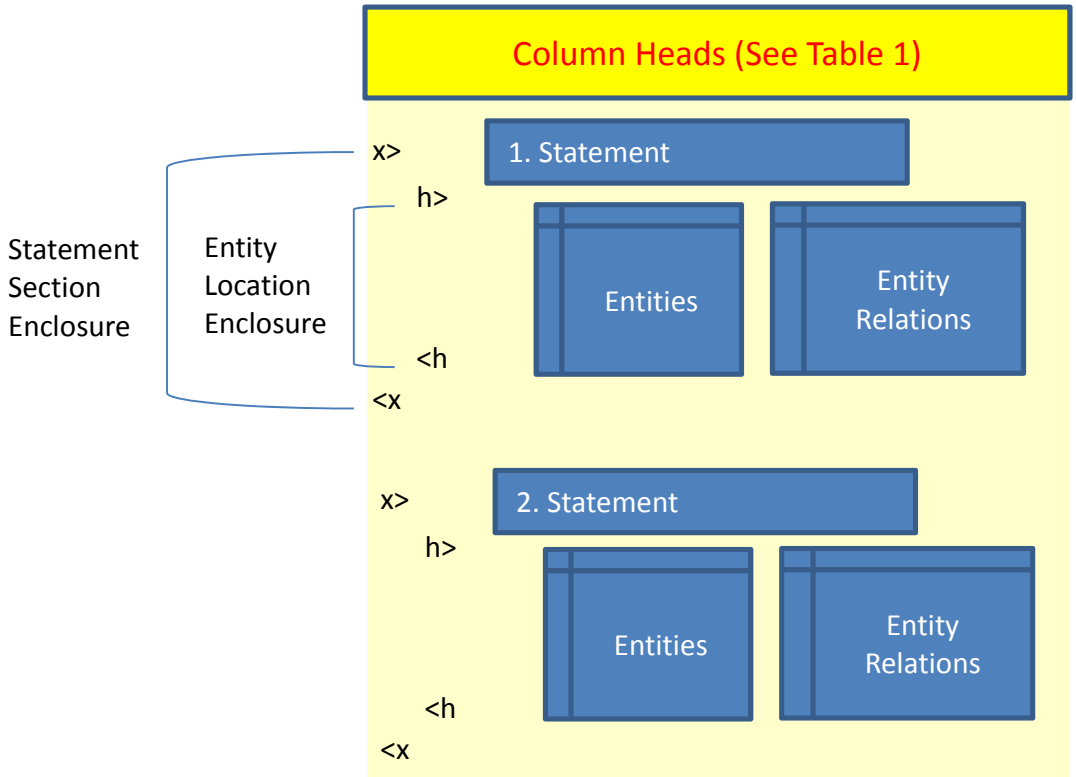


Table 1. Required Column Heads in a row

About Statements

#	Column	Labels	Content (Markers or symbols)
1	A	Attributes	Hide and show markers (x>,<x, h>,<h)
2	B	Qn	A question or statement number
3	C	Qt	Texts of a question or statement

About Entities

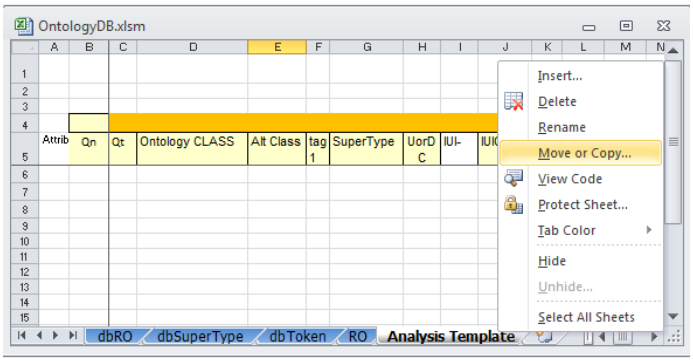
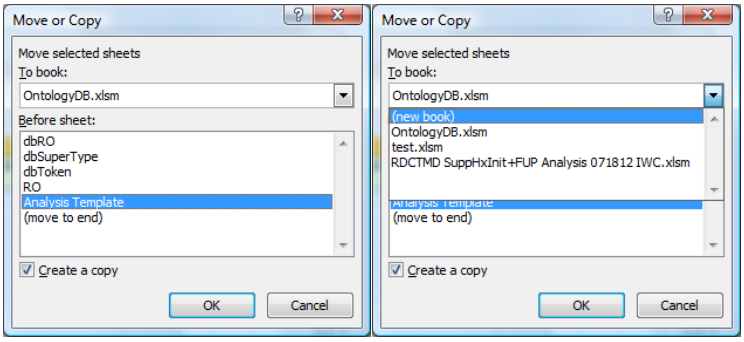
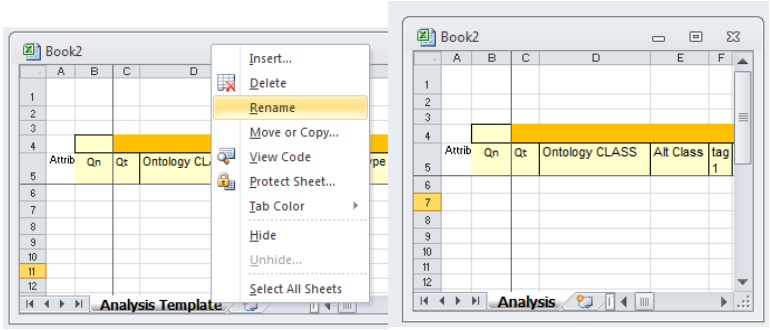
#	Column	Labels	Content
4	D	Ontology CLASS	Entity names
5	E	Alt Class	Alternative entity names
6	F	tag 1	inexplicit instance marker (ie)
7	G	SuperType	Supertype from SuperType Library
8	H	UorDC	Universal or Defined Class (U, DC)
9	I	IUI-	IUI number
10	J	IUICode	IUI code
11	K	ID-	ID number
12	L	IDCode	ID code
13	M	InstanceOrigin	the origin of a instance (a question number in column B)
14	N	tag 2	reused instance marker (re)
15	O	INSTANCE DESCRIPTION	Description about this instance
16	P	PSEUDO-FORMALIZATION	Pseudo-formalization of instance description
17	Q	tag 3	Customized marker

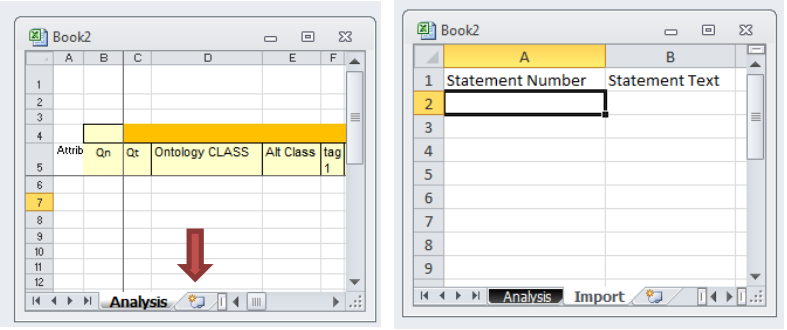
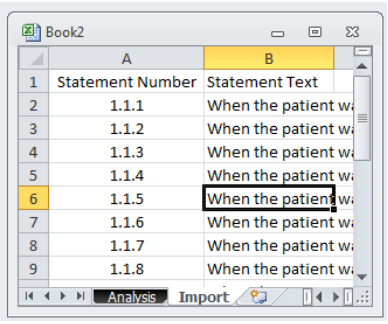
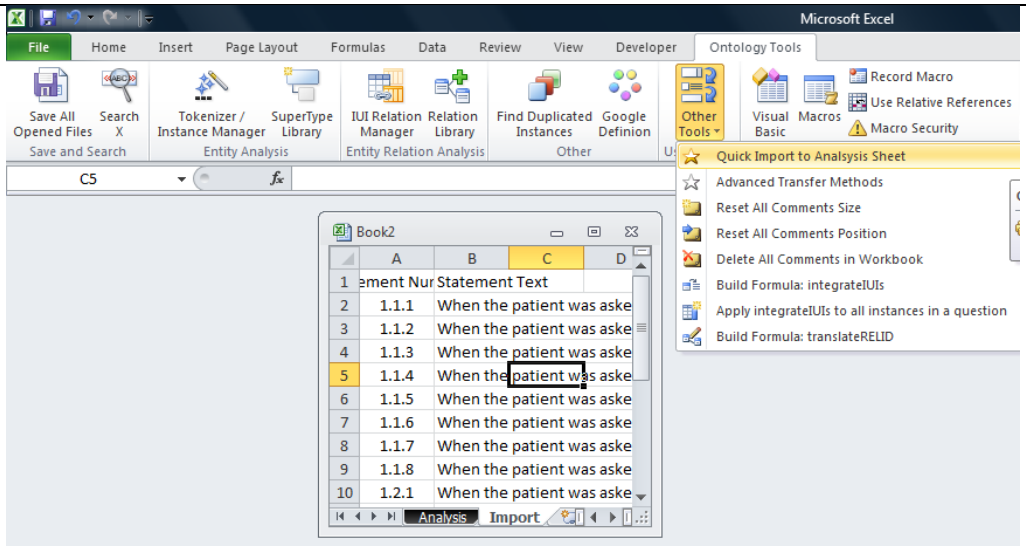
About Relations

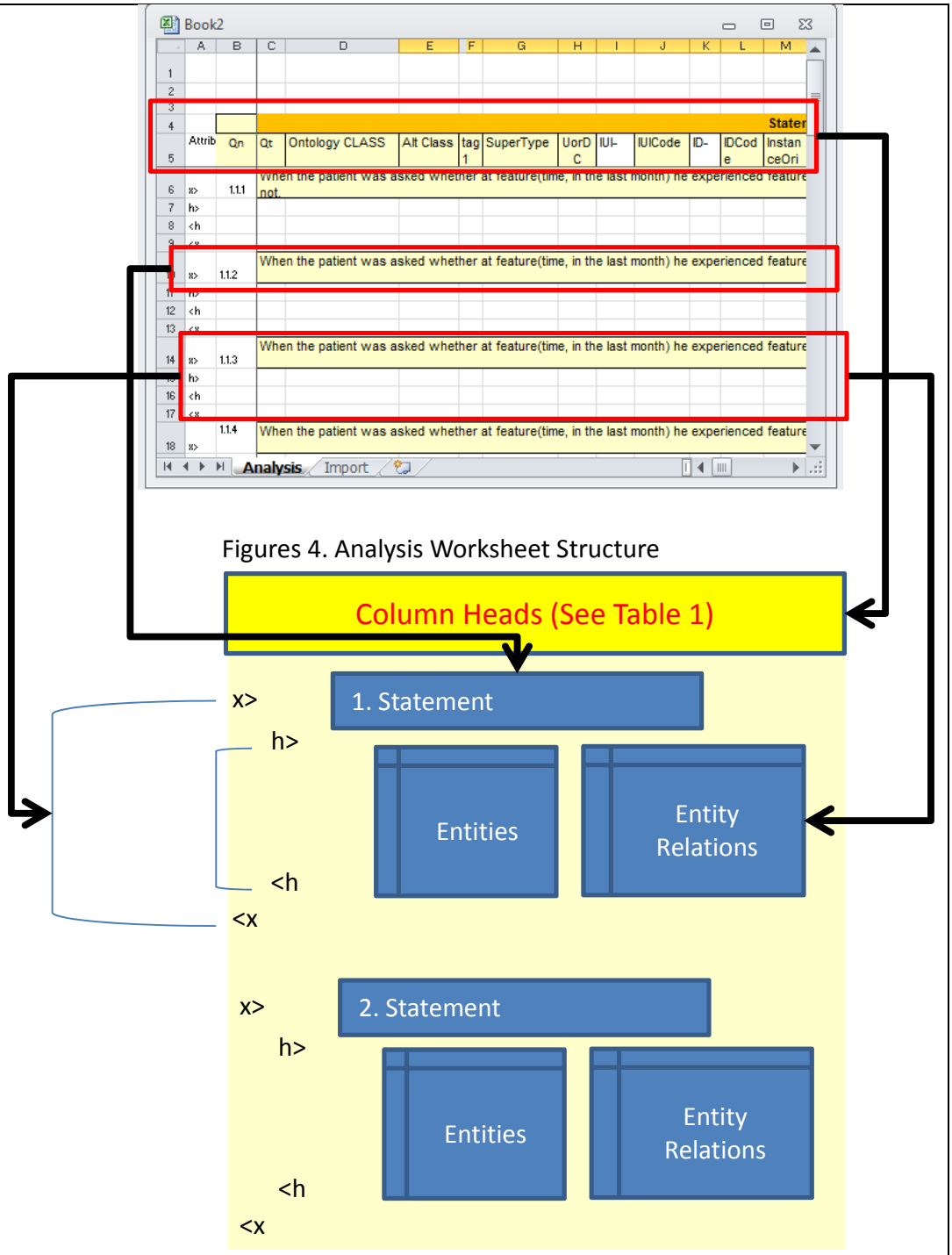
#	Column	Labels	Content
18	R	RelationDomain	IUI-code of this instance
19	S	RelID	Relation Id from Relation Library
20	T	RelationOntology	Relation Ontology from Relation Library
21	U	RelationRange	IUI-code of the Range instance
22	V	preposition	preposition of time instance
23	W	RelationTime	IUI-code of the time instance

How to Create Your Analysis File

Here is a step-by-step instruction on how you can convert your list of “single analyzable statements” into the format Ontology Tools can work with.

Steps	To Do	How
Step 1	Import an Analysis Sheet Temple from Ontology DB.	1. Open the “Ontology DB.xlsm” > **right click** sheet “Analysis Template” > choose “Move or Copy”  2. Select “Analysis Temple” in the list > Check the “Create a copy” option in lower left corner > Choose a file you want a copy of Analysis Temple goes to or create a new book (as shown).  3. Rename this new sheet as “Analysis”. 

Step 2	Arrange your single analyzable statement into a simple importable format	1. Insert a blank worksheet in the new workbook you just created with “Analysis” worksheet by click the tiny icon next the “Analysis” (indicated in red arrow) > and rename it to “Import” (Similar to step1.3.)  2. In the sheet “Import”, Copy and paste the statement numbers in to column A and statement full texts in column B 
Step 3	Run Ontology Tools > Other Tools > **Quick Import to Analysis Sheet** and follow the prompts	

	End Result of the import process	

Figures 4. Analysis Worksheet Structure

C. UserForms /VBA Modules

Userforms /VBA Modules in Ontology Tools are interconnected and serve different functions.

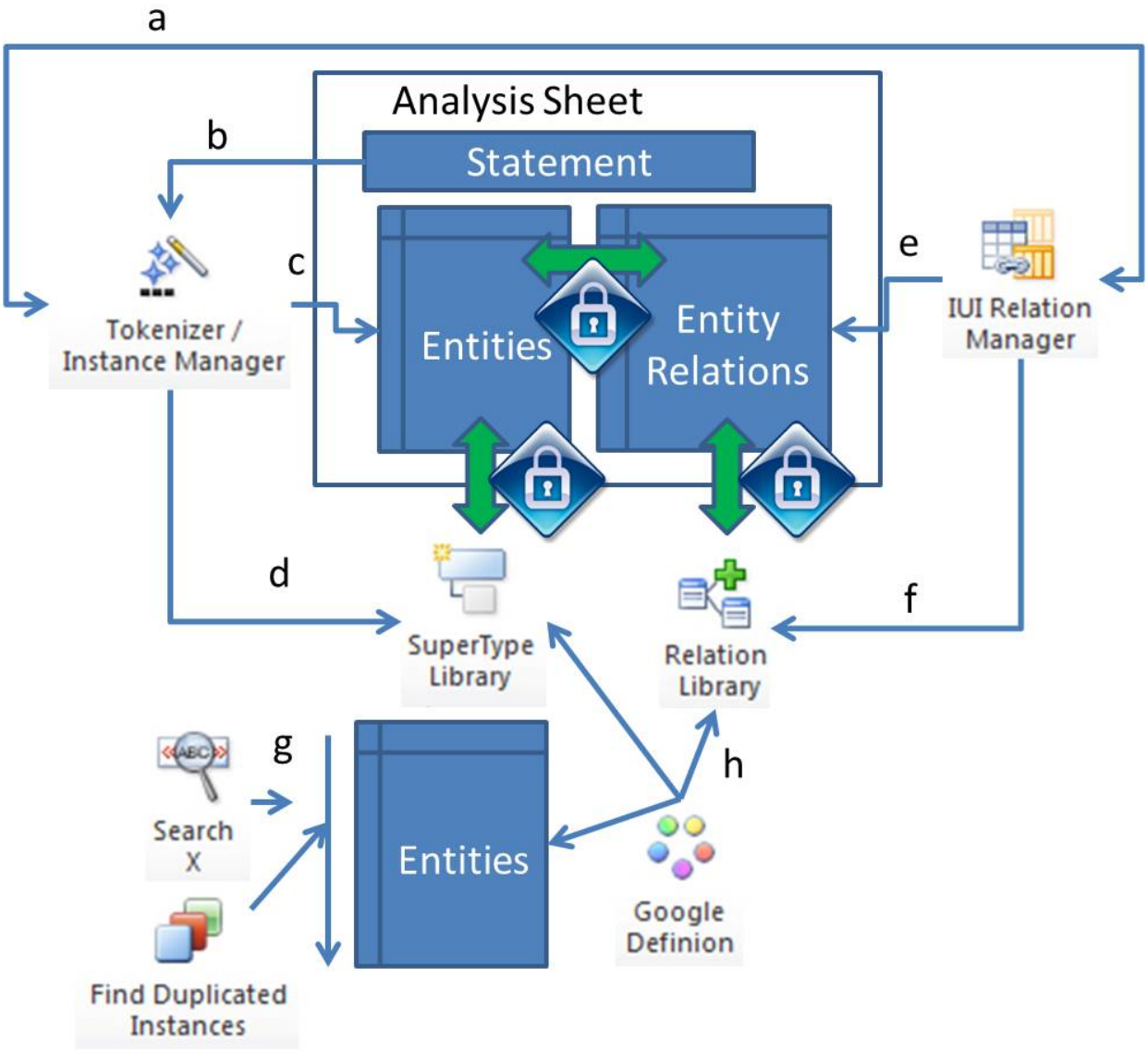


Figure 5. UserForms/ VBA Module Interact with Analysis Worksheet

a. Quick Switch between 2 apps
b. Tokenization: generate useful entities from statement
c. Insert, Delete, Modify Entities
d. Insert a direct formula link between SuperType Library and Entities
e. Insert, Delete, Modify Entity Relations
f. Insert a direct formula link between Relation library and relationship type
g. Search something or find a duplicates in a top-down fashion
h. Provide a "quick-but-dirty" definition when needed

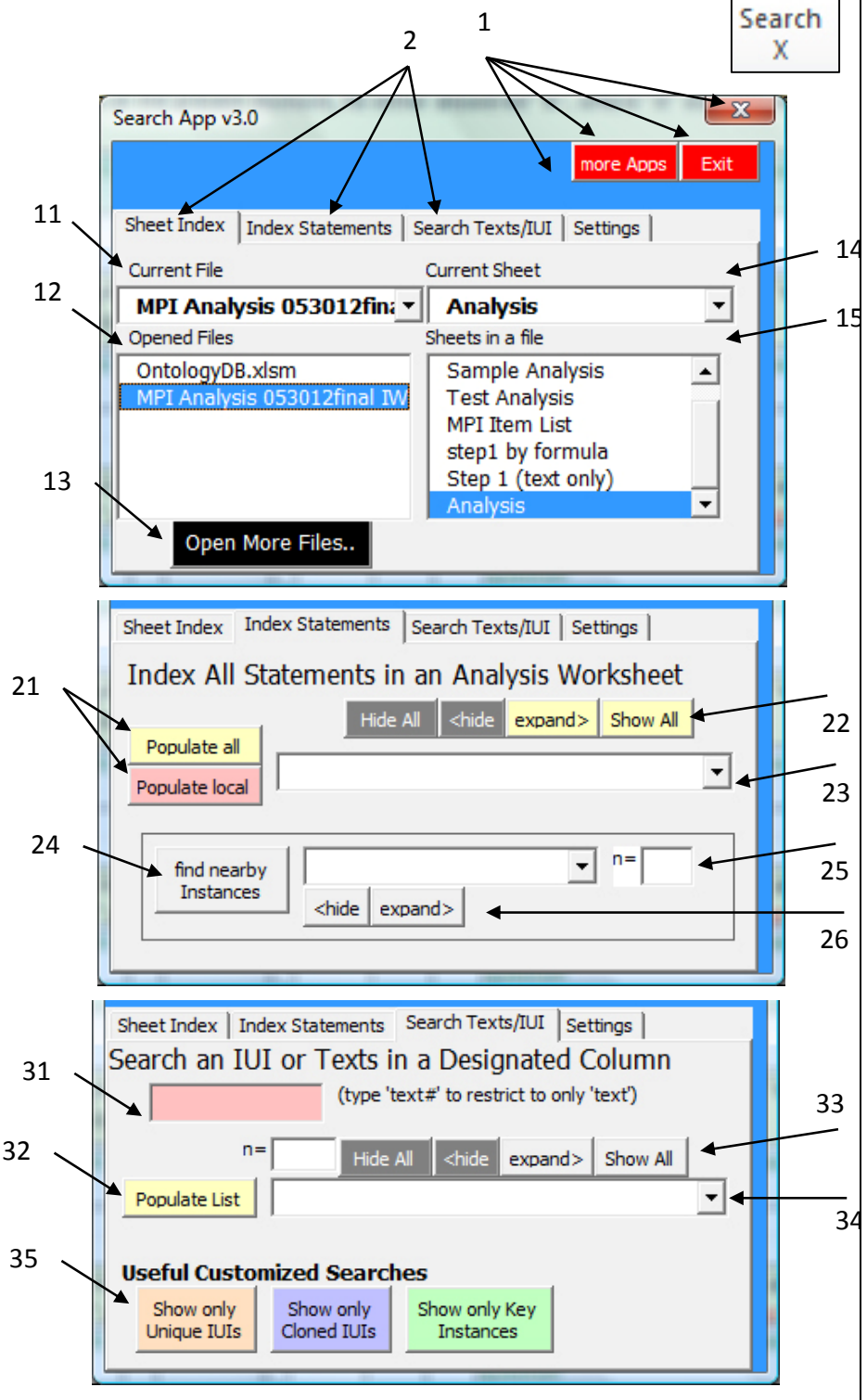
1. UserForms /VBA Modules > Save and Search > Search X

1. **General Controls**:
x = Window close button, disabled.
Exit or color background = hide this app. more Apps = open [App Central Station].
2. **Tabs**: Sheet Index = manage files and worksheets (11-15),
Index Statements = index all statement in the analysis sheet and allow to hide/show sections (21-26),
Search IUIs = search a specific text within a designed column and allow to hide/show sections (31-36).

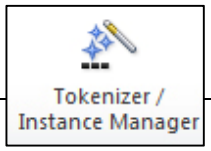
11. Current File
12. Opened Files =click to switch
13. Open more Files
14. Current Sheet =click to switch
15. Sheets in a file

21. Populate all or local = start searching all statement sections, or local pool (current position to the end).
22. Hide All , <hide / expand> , Show All = hide/ show buttons
23. Dropdown 1 = a populated index of statements
24. find nearby instances = find instance within a statement section
25 Dropdown 2 = a populated index of instances within a statement section
26 <hide / expand> = hide/ show selected instance

31. Search Box = search a text in a designated column
32. Populate List = start searching
33. Hide All , <hide / expand> , Show All = hide/ show buttons
34 Dropdown 3 (all rows with specified texts)
35. Useful Customized Searches



2. UserForms /VBA Modules > Entity Analysis > Tokenizer

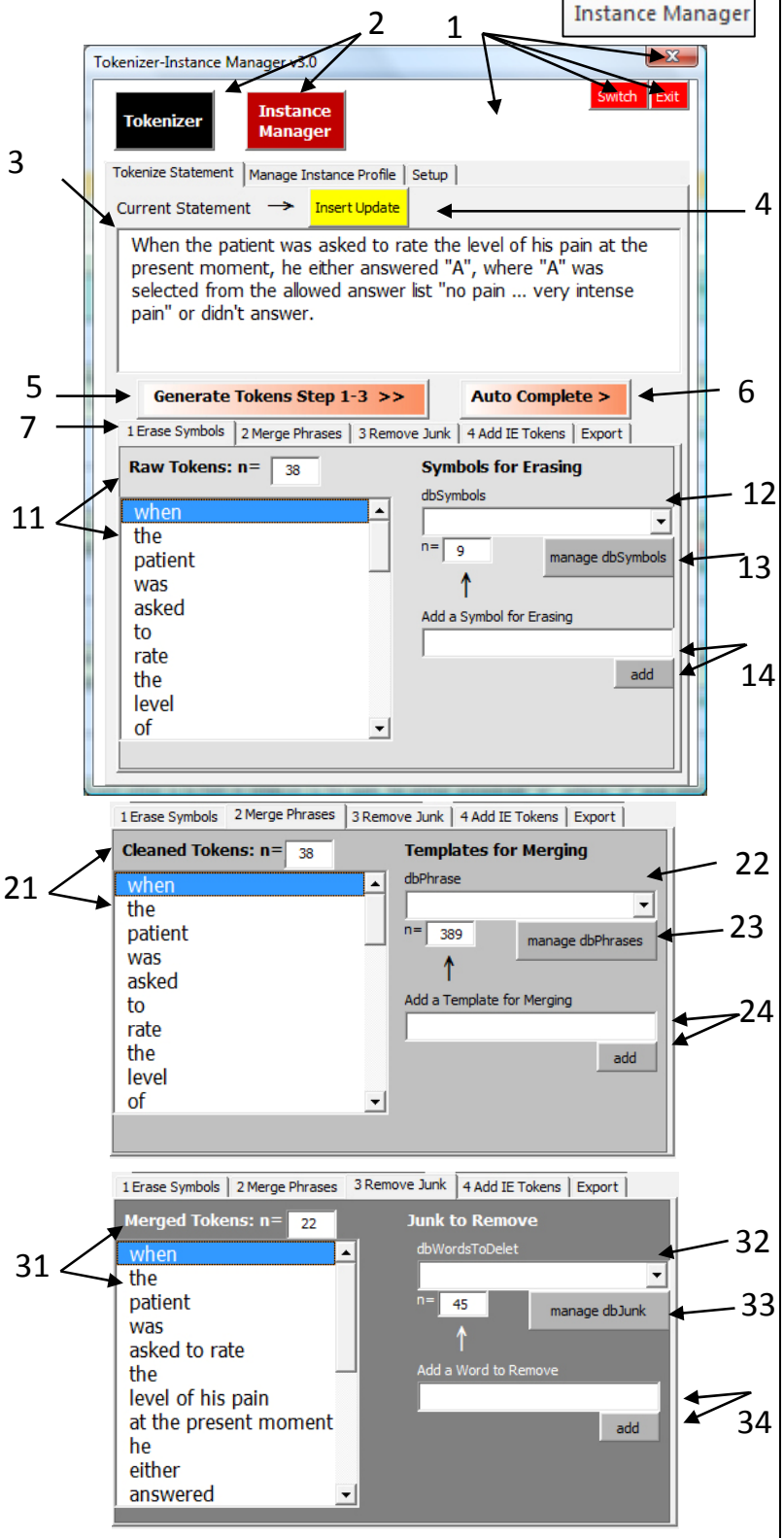


1. **General Controls**:
x = Window close button, disabled.
Exit = hide this app. Switch = open [App Central Station].
White Background = quick switch to IUI Relation Manager.
2. Tokenizer and Instance Manager (switch between 2 apps)
3. Current Statement.
4. Insert Update (insert changes made in 3.)
5. **Generate Tokens Step 1-3** = execute the steps of erasing symbols (step 1, 11-14), merging phrases (step 2, 21-24), removing junks (step 3, 31-34).
6. **Auto Complete** = automatically substitute key tokens with corresponding inexplicit tokens (step 4, 41-45), and prepare to export (export, 51-54)
7. Tabs: 1 Erase Symbols, 2 Merge Phrases, 3 Remove Junk, 4 Add IE Tokens, Export.

11. Raw Tokens = a collection of statement fragments or raw tokens.
12. dbSymbols = a collection of symbols that will be erased if found in 11.
13. manage dbSymbols = open location of dbSymbols
14. Add a symbol for erasing = add a symbol into dbSymbols

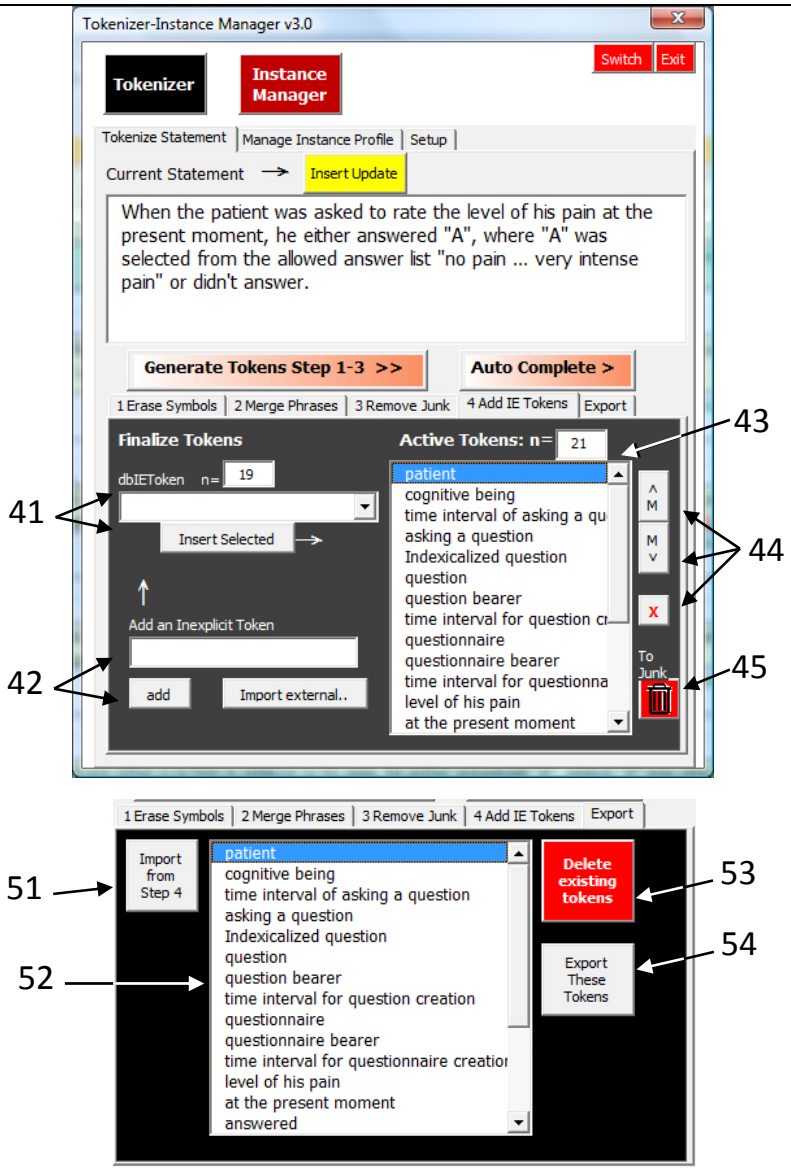
21. Cleaned Tokens = a collection of tokens contain no symbols.
22. dbPhrase = a collection of phrases that will be used as template to merge if matched sequence found in 21.
23. manage dbPhrase = open location of dbPhrase
24. Add a Template for Merging = add a word or phrase into dbPhrase

31. Merged Tokens = a collection of tokens contain some merged phrases.
32. dbWordsToDelete = a collection of words will be deleted if found in 31.
33. manage dbJunk = open location of dbJunk
34. Add a Word to Remove = add a word or dbJunk

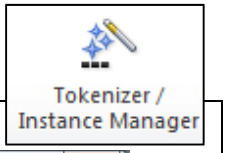


41. dbIEToken and Insert Selected = a collection Inexplicit Tokens and can inserted into 43.
42 Add an Inexplicit Token and Import external = add a words or phrase into dbIEToken by typing or Import from a worksheet.
43. Active Tokens = a collection of potentially useful tokens can be moved up or down in sequence or deleted by 44.
44. Controls to move the selected token in Active tokens list up or down in sequence, or delete from the list.
45. Move a selected token to 33.

51. Import from Step 4. (import active token list from Step 4 to Export page)
52. Tokens to export.
53. Delete existing tokens. (remove any items located in a designed area under the statement section in analysis worksheet)
54. Export these tokens. (insert 52 into the designed area under the statement section in analysis worksheet.)



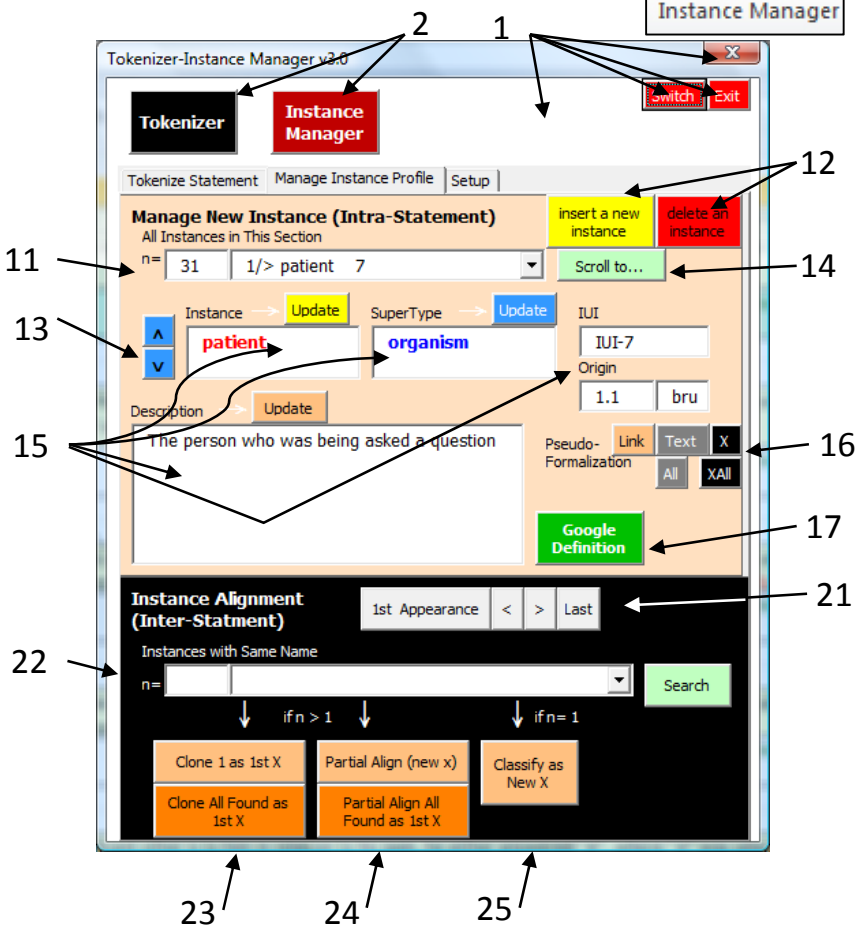
3. UserForms /VBA Modules > Entity Analysis > Instance Manager



1. **General Controls**:
x = Window close button, disabled.
Exit = hide this app. Switch = open [App Central Station].
White Background = quick switch to IUI Relation Manager.
2. Tokenizer and Instance Manager (switch between 2 apps)

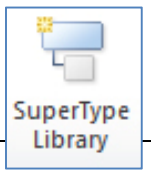
11. All Instances in this this section: Dropdown Box = a collection of instances in this statement section. (Arrow key up or down to navigate in the section)
12. Insert a new instance or delete an instance (insert a blank row in this section or delete the selected instance)
13 ^ and v = select the instance above or below the current position.
14. Scroll to = select an instance by using the mouse and Excel scroll bar.
15. Current Instance Information: Instance and Update , SuperType and Update (update by re-establishing link with dbSuperType), IUI, Origin, Description and Update.
16. Pseudo-Formalization (PF) : Link = PF using formula referenced to this section, Text = PF with texts only in this section, All =PF with text s only for current and all statement sections above, X = remove PF texts or formulas, XAll = remove current and all PF texts or formulas above.
17. Google Definition (Provide a "quick-but-dirty" definition if needed for 15.)

21. Quick-search similar instance nearby: 1st appearance, <, >, Last
22. Search and Index all similar instances located above current location: Search, Dropdown
23. Clone 1 as 1st X or Clone All found as 1st X = clone current instance identical to the 1st appearance of this instance, or clone current and all duplicated instances identical to the 1st appearance.



24. Partial Align or Partial Align All Found as 1st X = copy description and format of the 1st appearance of this instance but assign a new IUI-code, or do the same for all similar instance found above
25. Classify as New X = format this instance as Explicit or Inexplicit Instance.

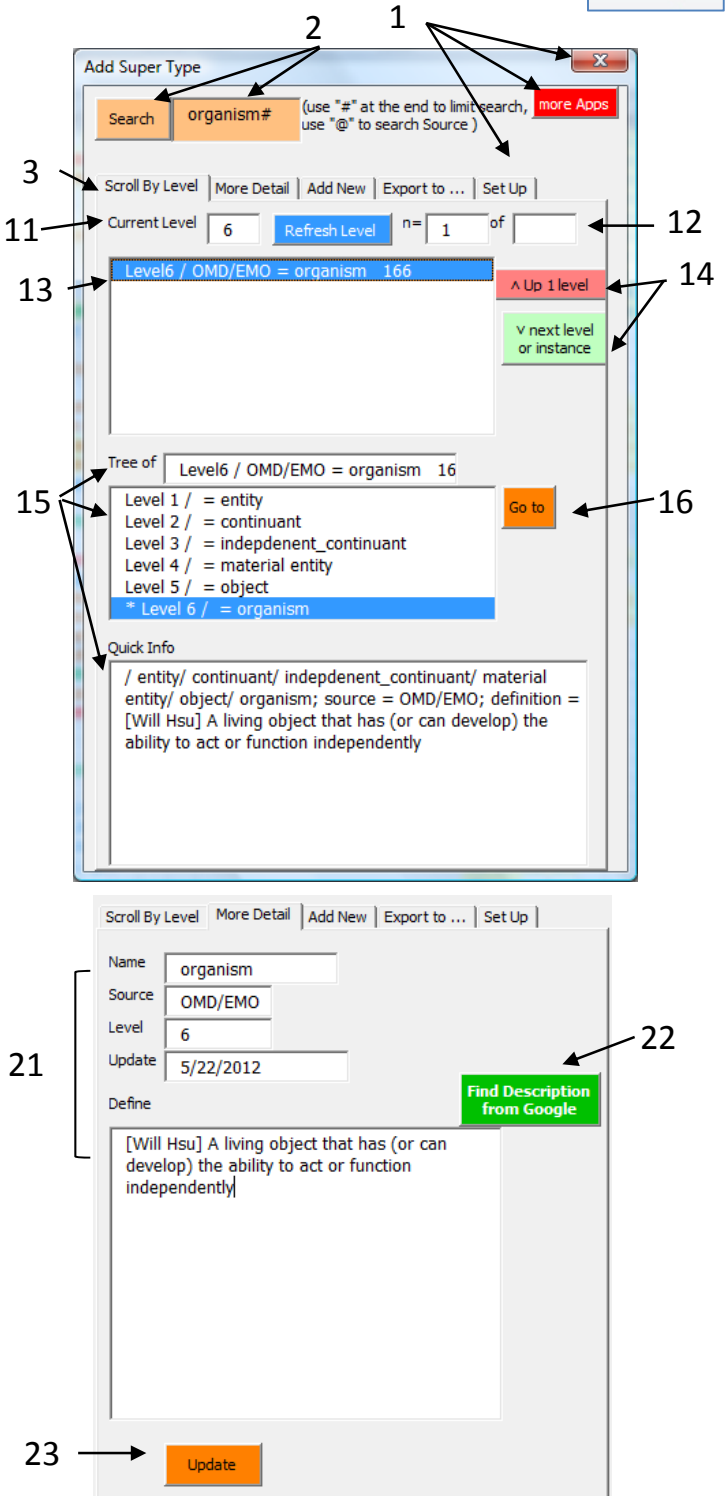
4. UserForms /VBA Modules > Entity Analysis > SuperType Library



1. **General Controls**:
x = Window close button, disabled. more Apps = open [App Central Station].
Blank Background = quick switch to Tokenizer/Instance Manager
2. Search an Supertype Category: Search Box, # to search for exact match, @ to search origin or source Reference Ontologies.
3. Tabs: Scroll by Level = Search Supertype by levels, More Detail = detail information of selected SuperType, Add New = a form enable users to insert new SuperType into the SuperType Library, Export to = export selected SuperType as a direct formula link

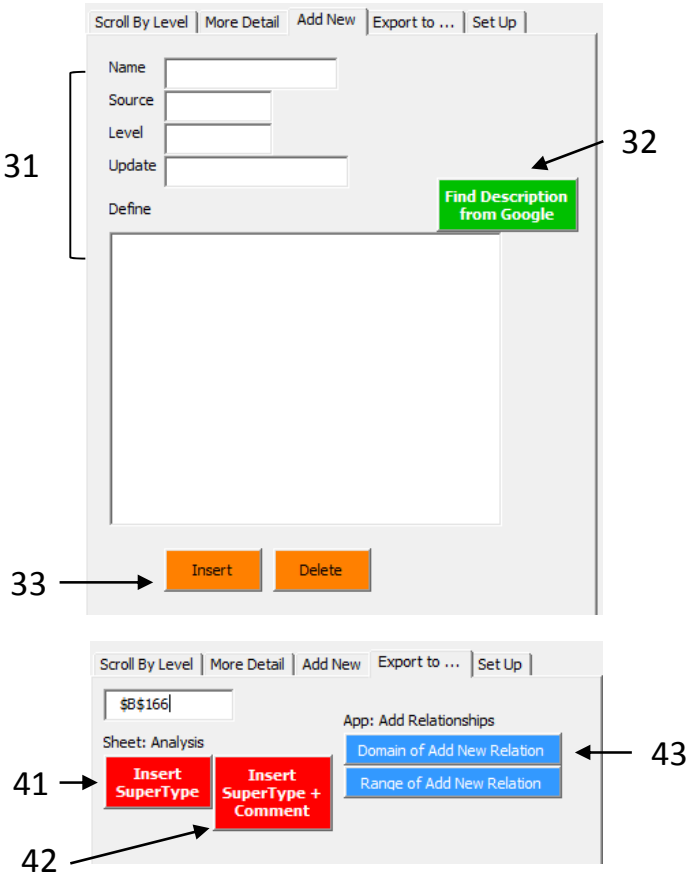
11. Current Level and Refresh Level: the hierarchy level the user is navigating in. The user can move up or down the hierarchy tree.
12. Supertype count in Current Level
13. SuperType List = a collection of SuperType categories in current level or matched search criteria.
14. ^ Up 1 level and ˅ next level or instance: The user can move up or down the hierarchy tree.
15. Selected SuperType = in 13, SuperType Tree = show hierarchy tree above, Quick Info = show tree, source, and definition of selected SuperType.
16. Go to = move up the tree in 15.

21. SuperType Information: Name, Source, Level, Update, Define.
22. Find Description from Google: (Provide a "quick-but-dirty" definition if needed for 21.)
23. Update = Insert changed information in 21 into SuperType Library.



31. SuperType Information: Name, Source, Level, Update, Define.
32. Find Description from Google: (Provide a "quick-but-dirty" definition if needed for 21.)
33. Insert or Delete= Insert a new SuperType in 31 into SuperType Library.

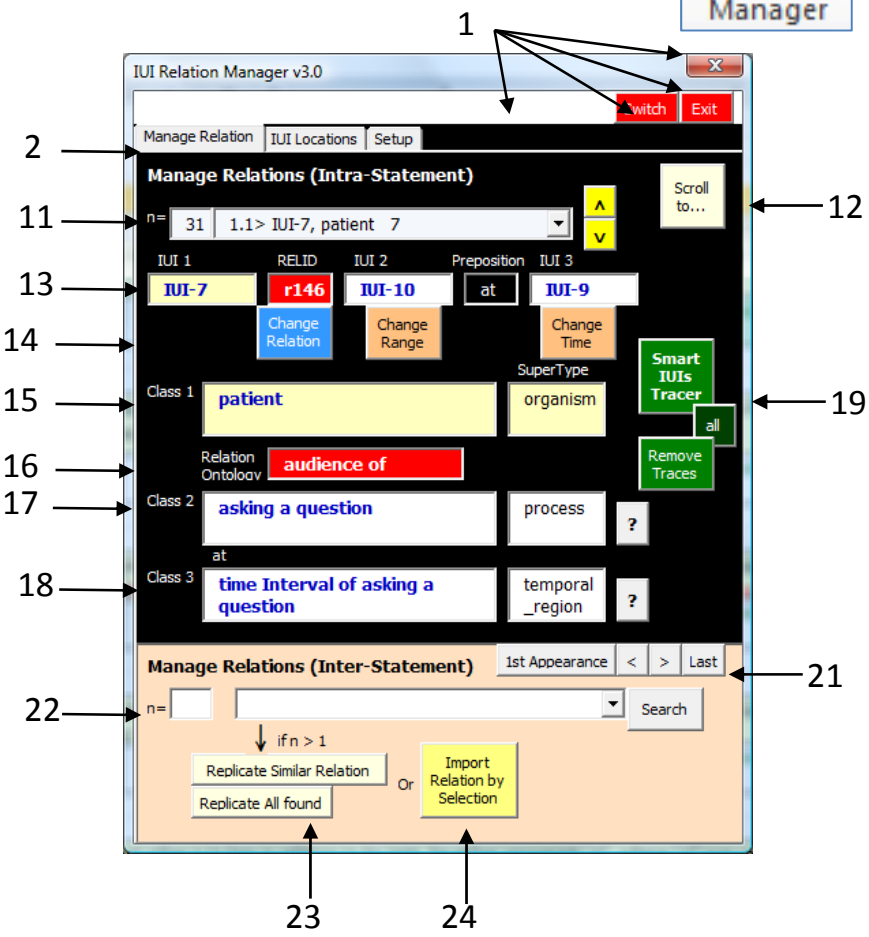
41. Insert SuperType = Insert a direct formula link of selected SuperType into Analysis Sheet
42. Insert SuperType + Comment: Insert a direct formula link of selected SuperType into Analysis Sheet with a Excel Comments containing detail information of the SuperType.
43. Insert SuperType into another App Relation Library



5. UserForms /VBA Modules > Entity Relation Analysis > IUI Relation Manager

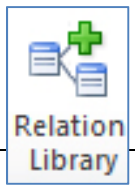
1. **General Controls**:
x = Window close button, disabled.
Exit = hide this app.
Switch = open [App Central Station].
White Background = quick switch to Tokenizer / Instance Manager.
2. Manage Relation. = a major tab in this app

11. All Instances in this this section:
Dropdown Box = a collection of instances in this statement section. (Arrow key up or down to navigate in the section).
^ and v = select the instance above or below the current position.
12. Scroll to = select an instance by using the mouse and Excel scroll bar.
13. 4 components of Relation:
IUI 1 = domain,
RELID,= Id of relation ontology in 16.
IUI 2 = Range,
IUI 3 = Time.
14. Change Relation = update RelID and 16 by re-establishing link with dbRO) , Change Range = change formula referencing location of IUI 2 and 17, Change Time = change formula referencing location of IUI 3 and 18.
15. Class 1: Translation of IUI 1/domain and its supertype.
16. Relation Ontology: Translation of RELID
17. Class 2: Translation of IUI 2/Range and its supertype. ? = Quick Info of Class 2.
18. Class 3: Translation of IUI 3/Time and its supertype. ? = Quick Info of Class 3.
19. Smart IUI Tracer = Trace and link to the 1st appearance of an IUI-Code,
All = Trace All Relations located in this Statement Section,
Remove Traces = Remove Traces by saving the file.



21. 1st appearance, <, >, Last = quick-search buttons, find a similar instance nearby.
22. Search and Dropdown = search and Index all similar relations located above current location:
23. Replicate Similar Relation or Replicate All Found = clone current relation identical to the 1st appearance of this relation, or clone current and all duplicated relations identical to the 1st appearance.
24. Import Relation by Selection = import a relation pattern by selecting with the mouse.

6. UserForms /VBA Modules > Entity Relation Analysis > Relation Library



1. General Controls:

x = Window close button, disabled.

more Apps = open [App Central Station].

Blank Background = quick switch to [IUI Relation Manager]

2. Search a Relation Category:

Search Box = # to search for exact match, @ to search origin or source Reference Ontologies.

3. Tabs:

Scroll by Level = Search Relation by levels.

More Detail = Detail information of selected Relation category in 13.

Add New Relation = A entry form enable users to insert new relation into the Relation Library,

Export to = Export selected relation as a direct formula link

11. Current Level = the hierarchy level the user is navigating in.

Refresh Level = The user can move up or down the hierarchy tree.

12. Relation count in current level.

13. Relation List = a collection of relation categories in current level or matching the search criteria.

14. ^ Up 1 level and ᵛ next level or instance = The user can move up or down the hierarchy tree.

15. Information on selected relation in 13:

Textbox = selected relation

Listbox = showing relation hierarchy

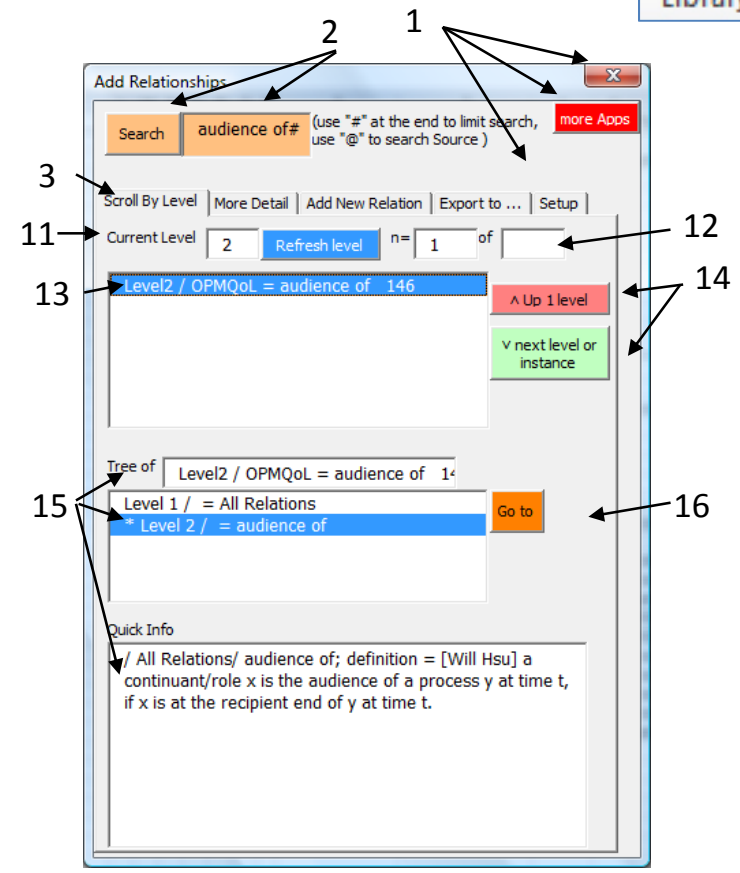
Quick Info = show relation hierarchy, source, and definition

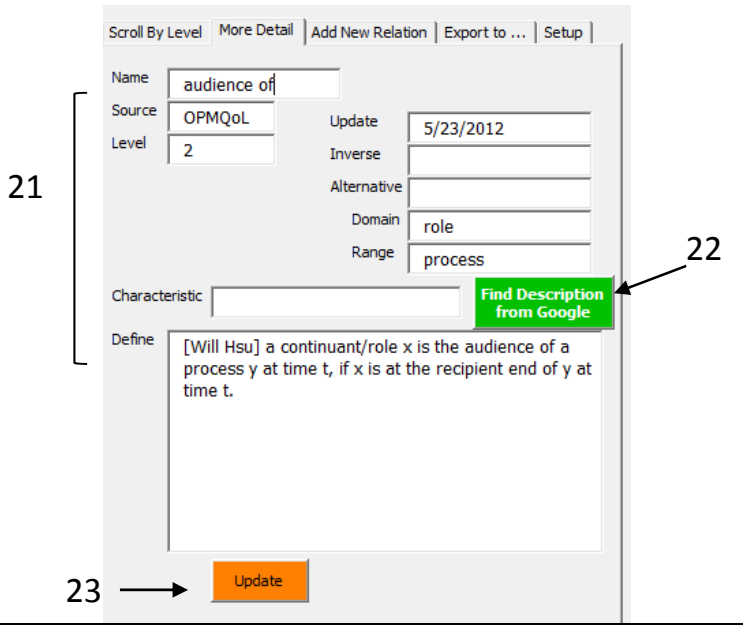
16. Go to = move up the hierarchy tree in 15.

21. Name, Source, Level, Update, Inverse, Alternative, Domain, Range, Characteristic, Define = Information of selected relation in 13.

22. Find Description from Google =Provide a "quick-but-dirty" definition if needed for 21.

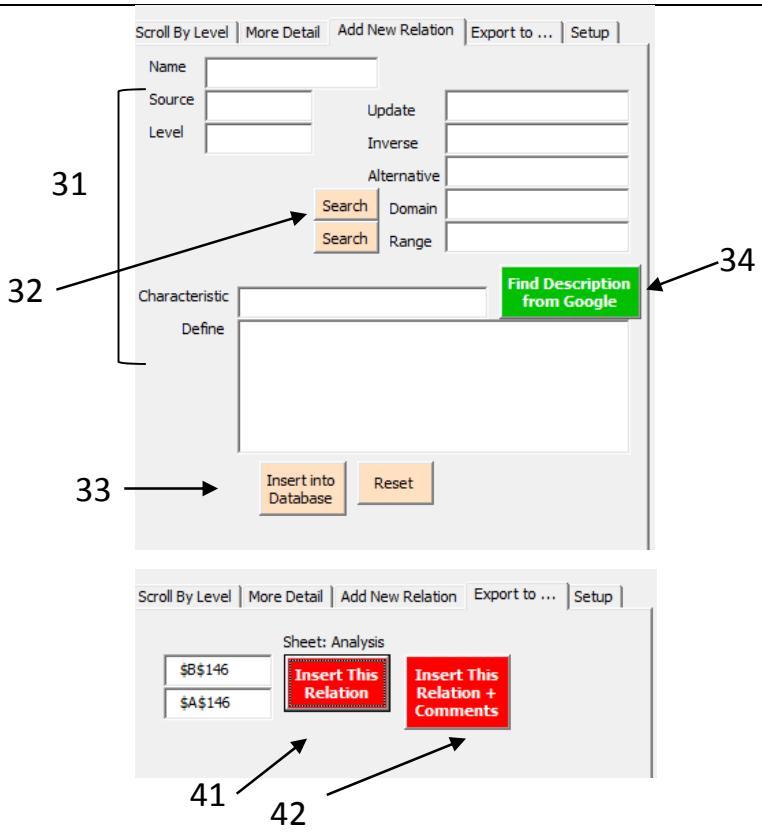
23. Update = Insert changed information in 21 into Relation Library.



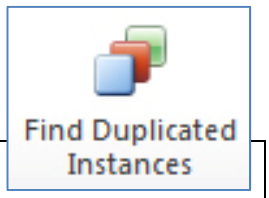


31. Name, Source, Level, Update, Inverse, Alternative, Domain, Range, Characteristic, Define = Information of selected relation in 13.
32. Search = find domain or range's supertype category from SuperType Library.
33. Find Description from Google = Provide a "quick-but-dirty" definition if needed for 31.
34. Insert into Database or Reset= Insert a new Relation in 31 into Relation Library.

41. Insert This Relation = Insert a direct formula link of selected Relation into Analysis Sheet
42. Insert This Relation + Comment: Insert a direct formula link of selected Relation into Analysis Sheet with an Excel comment containing detail information of the Relation.



7. UserForms /VBA Modules > Other > Find Duplicated Instances

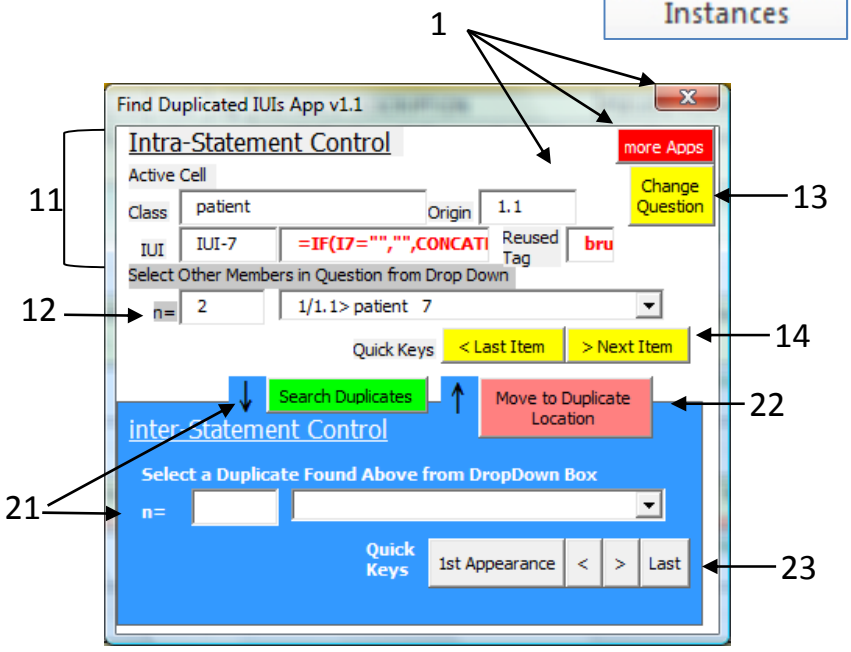


1. General Controls:

x = Window close button, disabled.
more Apps = open [App Central Station].
White Background = hide this app.

11. Class, Origin, IUI, Reused Tag = Information on current instance.
12. Dropdown Box = a collection of instances in this statement section (Arrow key up or down to navigate in the section).
13. Scroll to = select an instance by using the mouse and Excel scroll bar.
14. < last Item and >Next Item = Move to the instance above or below the current position

21. Search Duplicates and Dropdown = Search and Index all similar instances located above current location
22. Move to Duplicate Location = update Intra-Statement section when inter-statement section has changed
23. 1st appearance, <, >, Last = Quick-search buttons, locate similar instance nearby.



8. UserForms /VBA Modules > Other > Google Definition

1. General Controls:

x = Window close button, disabled.

White Background = hide this app.

2. Search Box = search a definition of a word

3. Result Text box = organized output from Google JSON Server, results can be copied and pasted into selected area