Ontologies as Nested Facet Systems for Human-Data Interaction[1]

Guo-Qiang Zhang [a,*], Shiqiang Tao [a], Ningzhou Zeng [b], and Licong Cui [a,**]
[a] *The University of Texas Health Science Center at Houston, Houston, Texas, USA*
E-mails: guo-qiang.zhang@uth.tmc.edu, shiqiang.tao@uth.tmc.edu, licong.cui@uth.tmc.edu
[b] *Department of Computer Science, University of Kentucky, Lexington, Kentucky, USA*
E-mail: ningzhou.zeng@uky.edu

Editors: Pascal Hitzler, Kansas State University, Manhattan, KS, USA; Krzysztof Janowicz, University of California, Santa Barbara, USA
Solicited reviews: Stefano Borgo, Consiglio Nazionale delle Ricerche (CNR), Italy; One anonymous reviewer

Abstract. Irrespective of data size and complexity, query and exploration tools for accessing data resources remain a central linkage for human-data interaction. A fundamental barrier in making query interfaces easier to use, ultimately as easy as online shopping, is the lack of faceted, interactive capabilities. We propose to repurpose existing ontologies by transforming them into nested facet systems (NFS) to support human-data interaction. Two basic issues need to be addressed for this to happen: one is that the structure and quality of ontologies need to be examined and elevated for the purpose of NFS; the second is that mappings from data-source specific metadata to a corresponding NFS need to be developed to support this new generation of NFS-enabled web-interfaces. The purpose of this paper is to introduce the concept of NFS and outline opportunities involved in using ontologies as NFS for querying and exploring data, especially in the biomedical domain.

Keywords: Web-interface, Ontology, Biomedical Big Data, Nested facet system, User experience

1. Introduction

When it comes to exploring and accessing biomedical data, often is the question asked: "Why can't it be as easy as shopping on Amazon?"

To answer this question, we need to identify the core technologies that made online-shopping experience "pleasant," and then hope to be able to apply a similar strategy for exploring and accessing biomedical data, big or small. Among many drivers of online-shopping [1], faceted search [2, 3] capability is perhaps one of the most ubiquitously applied information-retrieval techniques. Indeed, studies show that faceted search can help enhance user experience in a variety of settings [4–8].

Semantic labeling is the missing link between an entity (such as consumer goods for online shopping or study subjects in a clinical data warehouse) and ways to identify and accessing it through means such as a web-based user interface. This is well-articulated in a recent article by Balog [9] and in information organization as tags for folder and menu hierarchies [10–12].

Semantic labeling enables facets, such as *size, color, make, price* to be annotated for entities such as shoes in an online store. Faceted organization and presentation of metadata on products is the key mechanism that allowed consumers of web-sites to quickly narrow down from millions of products to items of interest using such simple facets. The entities for biomedical data, however, are highly complex and there does not exist a corresponding small set of semantic labels to support faceted search. For example, clinical data, captured as

[1]This work was supported in part by US National Cancer Institute under award R21CA231904 and by US National Science Foundation under awards IIS1931134 and ACI1626364.
[*]Corresponding author: guo-qiang.zhang@uth.tmc.edu.
[**]Co-corresponding author: licong.cui@uth.tmc.edu.

a part of patient care, are highly complex, and includes demographics, medical history, lab reports, diagnosis, medication, and discharge summaries.

Biomedical ontologies are suitable as semantic labels for biomedical entities. However, these ontologies, intended to model and capture concepts and their relations in the biomedical domain, are broad and complex. For example, SNOMED CT [13], the largest clinical terminology used worldwide, contains over 300,000 concepts and over 1.5 million relations. The National Cancer Institute thesaurus (NCIt) [14, 15], on the other hand, is a biomedical terminology managed by NCI Enterprise Vocabulary Services, containing more than 140,000 concepts related to cancer. Such size and complexity raise basic questions related to their potential role as facets for web-based user interfaces: What, if any, structural transformations are needed for ontologies to play the role of facets for information retrieval? Is it feasible to have ontologies to play the role of facets? What kind of desirable properties are required for ontologies to support facet-oriented user interaction? How to measure and evaluate the performance of this approach?

In this paper we propose the concept of *nested facet system* (NFS), outline a strategy to transform existing ontologies into NFS to support human-data interaction, and identify exemplar research questions related to the use of NFS to enhance user experience in human-data interaction. Unlike traditional faceted search, the intended users of interfaces supported by NFS are those equipped with some levels of knowledge in specific domains. Our motivation for NFS is to facilitate information retrieval in such specific domains, but NFS can also be readily implemented as a navigation interface for the corresponding underlying ontologies [16].

2. Nested Facet System

A facet is a semantic label of an entity along multiple possible axes or dimensions. Facets correspond to properties of the entity of interests. For example, online vendors use facets to label their product using readily available information about their type, brand, price, and support consumer shopping experience through faceted search [2].

A nested facet, or higher-order facet, is a facet that includes a (finite) collection of other facets as its components. In this context, traditional facets are primitive facets, those that are not made of other facets. A nested facet system is a set of nested facets (we call them facets from now on) with a taxonomy relation (i.e., subclass, subsumption, or hierarchical relation) among them.

Definition 1. *A nested facet system $\mathcal{F}$ is a finite set P (with its element called facet) and a collection of refinements $p \vdash \{q_1, \ldots, q_n\}$, with $p \in P$ (called the "head" of the refinement) and $q_i \in P$ for each $1 \leqslant i \leqslant n$ (called the "body" of the refinement), such that*

1. *Each element $p \in P$ is the head of at most one refinement;*
2. *The head of any refinement is not a part of the body of the same refinement.*

With respect to each refinement $p \vdash \{q_1, \ldots, q_n\}$, q_is are called sub-facets of p, and p is called a nested facet. Elements of P that do not have any sub-facets are called primitive facets.

The intuition for a refinement $p \vdash \{q_1, \ldots, q_n\}$ is that a complex facet p can be captured by a collection of sub-facets $q_1, \ldots, q_n$. Alternatively, if we think of p as a "query," then the logical disjunction of $q_1, \ldots, q_n$ is a "query expansion" for p.

Each NFS $\mathcal{F}$ induces a partial order in the following way. When $p \vdash \{q_1, \ldots, q_n\}$, we write $q_i \prec p$. We write $\preceq$ for the reflexive, transitive closure of $\prec$, which is a partial order on P (taking account for the equivalence class induced by $\prec$ when necessary).

To endow NFS' with their intended meaning, we treat facets as generalized semantic labels as follows. Given a set of entities E, a facet p with value space $\mathcal{D}(p)$ is a collection of parameterized semantic labels $p(t)$, such that for each member $e \in E$ and for each $t \in \mathcal{D}(p)$, e can be classified as having property $p(t)$ or not. For each $e \in E$, we write $e \models p(t)$ if entity e has facet p with value t. We write $[\![p(t)]\!]$ for the set $\{e \in E \mid e \models p(t)\}$ for t, and $[\![p]\!]$ for the set $\{e \in E \mid e \models p(t), t \in p(E)\}$. In extreme cases, we allow $\mathcal{D}(p)$ to be empty, and p can be specified without a parameter. For a refinement $p \vdash \{q_1, \ldots, q_n\}$, we write $p(\overrightarrow{t})$ for $\{q_1(t_1), \ldots, q_n(t_n)\}$, where $\overrightarrow{t} = (t_1, \ldots, t_n)$.

Definition 2. *When $[\![p]\!]$ is defined for each facet of an NFS $\mathcal{F}$, the triple $(E, \mathcal{D}, \models)$ is called an interpretation of $\mathcal{F}$. A refinement $p \vdash \{q_1, \ldots, q_n\}$ of $\mathcal{F}$ is sound with respect to an interpretation if $[\![q_i]\!] \subseteq [\![p]\!]$ for each $1 \leqslant i \leqslant n$. An NFS $\mathcal{F}$ is sound with respect to an interpretation if each of the NFS' refinement is sound.*

Proposition 1. *If $(E, \mathcal{D}, \models)$ is a sound interpretation for $\mathcal{F}$, then we have $[\![q]\!] \subseteq [\![p]\!]$ whenever $q \preceq p$.*

Definition 3. *We call $\mathcal{F}$ complete with respect to $(E, \mathcal{D}, \models)$, when it is the case that with respect to any refinement $p \vdash \{q_1, \ldots, q_n\}$ in $\mathcal{F}$, we have the property that for any $e \in E$, if $e \models p$, then for some i, $e \models q_i$ with $1 \leqslant i \leqslant n$.*

Proposition 2. *If $(E, \mathcal{D}, \models)$ is a complete interpretation for $\mathcal{F}$, then we have*

$$[\![p]\!] = \bigcup_{1 \leqslant i \leqslant n} [\![q_i]\!]$$

for each refinement $p \vdash \{q_1, \ldots, q_n\}$ in $\mathcal{F}$.

Note that the notation $\vdash$ is deliberatively suggestive of a potential connection with "Information Systems" [17, 18], part of domain theory [19] as a mathematical foundation for programming languages [20]. There appears to be potential formal connection to the notion of disjunctive information systems [18, 21].

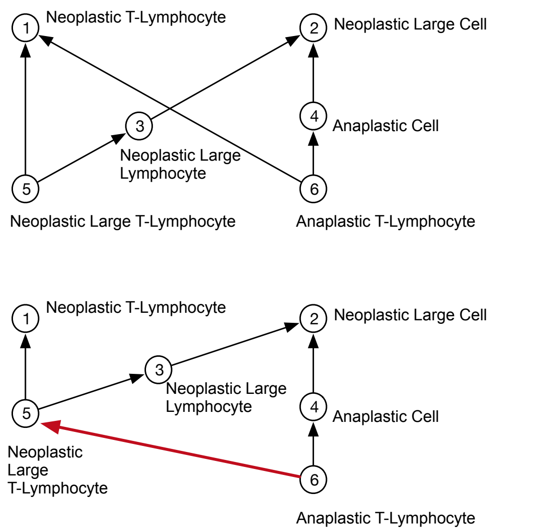

Fig. 1. Two example NCI Thesaurus fragments. Above: a fragment containing a bug. Below: fragment with the bug fixed by redirecting node 5 as a direct parent of node 6 (red edge).

3. Ontologies as Nested Facet Systems

Biomedical ontologies serve as the semantic scaffolding for us to fully capitalize on the transformative opportunities of the increasingly large amounts of digital data produced by the biomedical research enterprise. For example, BioPortal [22], the world's most comprehensive repository, contains over 600 ontologies and over 7 billion concepts that have been used to support a wide spectrum of scientific projects. Biomedical ontologies provide the basis for scientific rigor during the process of data collection, annotation, management, analysis, and sharing in biomedicine. They not only serve as metadata standards, but also play a vital role in down-stream systems as a declarative knowledge source [23]. For example, SNOMED CT [13], the most comprehensive and precise clinical health terminology product in the world, facilitates the clear exchange of health information in Electronic Health Records (EHRs), leading to higher quality, consistency and safety in healthcare delivery [24, 25].

Ontological systems are not designed *a priori* as nested facet systems. But what if we attempt to reuse them as facets to support user interfaces? An intuitive idea is to leverage the hierarchical or is-a relation, the structural backbone of most ontologies and simply treat *Ontological Concepts as Facets*.

For a given ontology such as SNOMED CT, we can treat each concept c as a facet p, and build a nested facet system by letting $p \vdash \{q_1, \ldots, q_n\}$ if the concepts corresponding to the q_i's are the (immediate) lower neighbors of p. In other words, if p is the facet corresponding to c, and q_i's are the facets corresponding to all the (immediate) lower neighbors of c with respect to the hierarchical relation, then make p a nested facet with q_is its components.

For this (very reasonable) intuition to work, the following questions must be answered:

1. Does this construction obey the soundness property mentioned at the end of the previous section?
2. Does this construction obey the completeness property, mentioned at the end of the previous section?

Intuitively, soundness means that all items below each facet are relevant to the facet. Completeness means that any items or facets relevant to a specific facet are already contained in and accessible through the facet. The soundness and completeness properties of NFS directly affect query performance in terms of precision and recall. Incomplete facets will reduce recall, while unsound facets will reduce precision. Top of Figure 1 contains an incomplete facet, in that concept node 5 as a facet missed the sub-facet represented by concept node 6.

Interestingly, similar properties of soundness and completeness have been studied in the area called Ontology Quality Research (OQR [26]) encompassing ontology quality auditing, assurance, and evaluation [27, 28]. For example, OQR method can identify a missing is-a relation (incompleteness) in the top fragment of Figure 1 and automatically suggest the addition of is-a in the lower part of Figure 1. The addition of this is-a edge (in red) makes the facet represented by node 5 "more complete" because it now includes node 6 as a sub-facet (as it should be). The goal of OQR is to develop methods and tools to detect [29, 30], identify [31], and address [32–34] quality issues in ontologies. This is a particularly important area in the biomedical domain, because of the significance, scope, complexity, manual involvement and evolving nature of biomedical ontologies that are intended to serve as terminology standards, as well as to codify knowledge at the same time.

Identifying quality issues in ontologies such as unsoundness or incompleteness is a task similar to finding bugs in software. Just as there is no single "recipe" to catch and fix all software bugs, no single method is expected to exist that addresses all ontology quality issues all at once. Similarly, for NFS, a single method to ensure and allow us to formally prove its soundness and completeness is unlikely. Instead, we see the development of methods to "improve" soundness and completeness of NFS' derived from ontological systems, leading to meaningful enhancement of the performance of NFS for information retrieval tasks.

In the following sections we discuss such questions in more depth using biomedical ontologies and clinical data resources as examples, and provide use cases to demonstrate the feasibility and work involved to implement this approach.

4. Data Resources and Related Ontologies

An array of biomedical datasets in the context of human health exists but there is a general lack of faceted interfaces to facilitate data exploration and information retrieval. In most of the cases, ontological systems have already been used for annotating or labeling the backend data but their interface roles have not been fully exploited. This state of affairs represents a ripe and rich setting for developing and implementing NFS to facilitate cohort discovery and sub-group analysis. This section provides a brief synopsis of these data resources and the associated ontological systems as an illustration of a targeted application area for NFS.

4.1. Clinical Data Warehouse

The entity E for clinical data consists of patients. Clinical data from EHRs are critical for analyses to improve health care delivery. Clinical data warehouses are EHR data made available for research. Examples include i2b2 data warehouses [35, 36], PCORnet – the National Patient-Centered Clinical Research Network [37], and Observational Health Data Sciences and Informatics (OHDSI) research network [38] with an open, community data standard called the Observational Medical Outcomes Partnership (OMOP) Common Data Model. SNOMED CT is a common ontological component of all these data sources.

4.2. Health Claims Data

Health claims data (also called administrative data) such as Cerner Health Facts, IBM Market Analytics, and Optum Health Data and Analytics, are those collected for the purpose of health insurance claims. They include information at the patient encounter level regarding diagnoses, treatments and billed and paid amounts. This is a valuable data source for research aimed at driving improvements in population health to address issues related to cost, quality and outcomes. The use of administrative data can complement EHR data by providing a regional or national scale view. Because of the health claims context, main vocabularies for health claims data involve diagnosis (ICD 9 and ICD 10), procedure code (CPT), and medication (RxNorm).

Clinical data and health claims data are domain-agnostic: they cover the entire spectrum of disorders and disease domains. Domain-specific data resources, however, are those cover a signal medical specialty, but with greater depth. We highlight several such resources next.

4.3. The National Sleep Research Resource - NSRR

The gold standard for sleep diagnosis is polysomnography (PSG), which monitors physiological processes including electroencephalogram (EEG - brain waves), electromyogram (EMG - muscle tone), and electro-occulogram (EOG - eye movements). The recorded polysomnograms provide comprehensive data about biophysical changes that occur during sleep and characterize the association between sleep and other public health related problems. The NSRR [39, 40] is a retrospectively annotated repository of 30,000 overnight sleep recordings. The NSRR offers free and open web access to large collections of de-identified, well-annotated national repository of sleep data, including PSGs which are linked to risk factor and outcome data for participants in major NIH studies. Since its launching in 2014, 282TB of data have been shared by over 3,000 users around the world through the NSRR portal sleepdata.org.

NSRR uses the Sleep Domain Ontology [41] as the canonical vocabulary for across-study data mapping.

4.4. The Center for SUDEP Research - CSR

The Center for Sudden Unexpected Death in Epilepsy (SUDEP) Research [42] manages another domain-specific clinical research data resource. The CSR has prospectively collected high grade multimodal data including high-resolution electroencephalographic signal, research-grade brain MRI, biochemical and DNA samples together with detailed phenotypic data for more than 3,000 epilepsy patients. Similar to NSRR, a disease-specific ontology called Epilepsy and Seizure Ontology [43] has been created as a part of the CSR informatics infrastructure process.

4.5. Cancer Registries

For cancer research, the US National Cancer Institute's Surveillance Epidemiology and End Results (SEER) program [44] coordinates a collection of state-based SEER registries. These state-centered cancer registry receiving data about new cancer cases from healthcare facilities and physicians within the state. Typically, five aspects of data are captured: patient data, case data, follow-up, therapy data and pathology reports. Patient data consists of variables including various patient-related information such as demographics, race, ethnicity, smoking, and clinical trial participation information. Case data captures variables for diagnosis, morphology, staging, biomarkers, and other categories. Follow up information contains variables including follow-up physician, date of last contact, survival status, and cancer status. Therapy data records variables with information on surgery, chemotherapy, radiation, and other treatment modalities.

In general, SEER data are considered to be among the most accurate and complete population-based cancer registries in the world that includes stage of cancer at the time of diagnosis and patient survival data. Cancer registries uses NAACCR data dictionary [45] for variable definition, and is only partially mapped to NCIt. This is where work on primitive facets is needed in order to use NCIt as NFS.

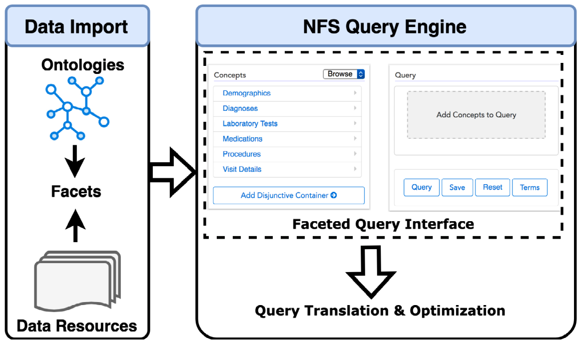

Fig. 2. High level functional architecture of an NFS-based system.

5. Implementation Strategy

The following steps are typically involved in developing an NFS-based query engine for a data source (see Figure 2 for a functional architecture).

1. Identify or develop a domain ontology covering the conceptual scope of the data source. If multiple ontologies are used, ontology merging would be a necessary step involved in developing such a domain ontology.
2. Construct a mapping from the data dictionary for the data source to concept of the domain ontology.
3. Convert the domain ontology to NFS and implement NFS-based query interface by systematically extracting the "refinement" structure of nested facets from the hierarchical relationships of the ontology following the method given in Section 3.
4. Implement an appropriate query optimization strategy dedicated to the data source as a database. Transformation to a NoSQL database such as MongoDB may be desirable depending on the data source.

Model-View-Controller [46], a well-established and popular web-based application development paradigm, is a suitable approach for developing an NFS-based system, particularly for the clinical informatics domain [47].

6. Opportunities and Challenges

For disease-specific domains such as sleep and epilepsy, we have developed NFS query interfaces such as x-search [48] and Multi-Modality Epilepsy Data Capture and Integration System (MEDCIS [49]).

x-search is a cross-cohort query and exploration system to enable researchers to query patient cohort counts across a growing number of completed, NIH funded studies in the NSRR. x-search is public available at `https://x-search.net` covering over 26,000 unique subjects. The canonical data dictionary, Sleep Domain Ontology [41], covers over 900 common data elements across a dozen cohort studies in NSRR. x-search has received over 2,300 queries by users from 16 countries since its initial launch [48]. For epilepsy, the MEDCIS interface uses a dedicated Epilepsy and Seizure Ontology [43] to drive an NFS-based query interface. MEDCIS is the main query interface for CSR data, integrating curated multi-modality clinical data of 2,000 epilepsy patients from 8 medical centers.

Based on our experience, benefits of an NFS-based query interface include:

1. It provides an intuitive interface for users to navigate to a specific concept of interest and specify the corresponding query criterion in a menu-driven, templated style.
2. The same boolean query can be constructed in a more efficient manner, usually involving only half of the time than that is needed for alternative interfaces without involving NFS.
3. A query optimization strategy can be readily implemented by precomputing queries corresponding to primitive facets and ordering the query execution sequence based on the result sizes for primitive facets.

Such benefits have been studied in the clinical data warehouse setting [50] but we also encountered challenges that seem to be typical in developing an NFS-based query interface:

1. There is no clear and efficient way to guarantee the soundness and completeness properties of NFS in general. For example, even though SNOMED CT and NCIt satisfy the soundness and completeness properties "for the most part" using the NFS refinements specified in Section 3, enough facet instances exist where such properties are violated [29]. Such violations affect the soundness and completeness properties of facets, leading to reduced precision and recall for query interfaces using NFS. Interestingly, non-lattice auditing methods can precisely identify and potentially fix such issues [30–34].
2. Primitive facets are not always specified and ready for use. For example, for Cancer Registries, the common data dictionary exists (i.e. NAACCR), but not all of its variables have been structurally mapped to appropriate NCIt terms both in value type and value range. Effort is needed to construct such a mapping (once only, though) before data dictionary variables can be used as primitive facets.
3. When a domain ontology is large and deep (e.g. SNOMED CT), interface response can be sluggish if the hierarchical (sub-facet) interface widget rendering algorithm is not optimized.

7. Conclusion

We outlined a general approach for constructing nested facet systems from ontologies. We highlighted use cases for clinical data, and discussed progress and remaining challenges. Given the importance of faceted search, our proposed approach deserves further study. Efforts in developing experimental interfaces supporting NFS will be highly desirable and impactful for accessing biomedical data for research.

References

[1] T.P.Y. Monsuwé, B.G. Dellaert and K. De Ruyter, What drives consumers to shop online: A literature review. International Journal of Service Industry Management, in: *Information Systems Res*, Citeseer, 2004, doi:10.1108/09564230410523358.

[2] D. Tunkelang, Faceted search, *Synthesis lectures on information concepts, retrieval, and services* **1**(1) (2009), 1–80, doi:10.2200/S00190ED1V01Y200904ICR005.

[3] J. Koren, Y. Zhang and X. Liu, Personalized interactive faceted search, in: *Proceedings of the 17th international conference on World Wide Web*, ACM, 2008, pp. 477–486, doi:10.1145/1367497.1367562.

[4] J.C. Fagan, Usability studies of faceted browsing: A literature review, *Information Technology and Libraries* **29**(2) (2010), 58–66, doi:10.6017/ital.v29i2.3144.

[5] M.A. Hearst, UIs for faceted navigation: Recent advances and remaining open problems, in: *HCIR 2008: Proceedings of the Second Workshop on Human-Computer Interaction and Information Retrieval*, 2008, pp. 13–17, doi:10.1.1.143.5111.

[6] L. Cui, R. Carter and G.-Q. Zhang, Evaluation of a novel conjunctive exploratory navigation interface for consumer health information: a crowdsourced comparative study, *Journal of medical Internet research* **16**(2) (2014), e45, doi:10.2196/jmir.3111.

[7] G.-Q. Zhang, T. Siegler, P. Saxman, N. Sandberg, R. Mueller, N. Johnson, D. Hunscher and S. Arabandi, VISAGE: a query interface for clinical research, *Summit on Translational Bioinformatics* **2010** (2010), 76.

[8] X. Niu, X. Fan and T. Zhang, Understanding Faceted Search from Data Science and Human Factor Perspectives, *ACM Transactions on Information Systems (TOIS)* **37**(2) (2019), 14, doi:10.1145/3284101.

[9] K. Balog, Semantically Enriched Models for Entity Ranking, in: *Entity-Oriented Search*, Springer, 2018, pp. 101–143, doi:10.1007/978-3-319-93935-3_4.

[10] O. Bergman, N. Gradovitch, J. Bar-Ilan and R. Beyth-Marom, Folder versus tag preference in personal information management, *Journal of the American Society for Information Science and Technology* **64**(10) (2013), 1995–2012, doi:10.1002/asi.22906.

[11] G.-Q. Zhang, G. Shen, Y. Tian and J. Sun, Concept analysis as a formal method for menu design, in: *International Workshop on Design, Specification, and Verification of Interactive Systems*, Springer, 2005, pp. 173–187, doi:10.1007/11752707_15.

[12] G.-Q. Zhang and Y. Tian, ACOSys: An Experimental System for Automated Content Organization, *ICCS 2005* (2005), 186.

[13] K. Donnelly, SNOMED-CT: The advanced terminology and coding system for eHealth, *Studies in health technology and informatics* **121** (2006), 279.

[14] S. De Coronado, M.W. Haber, N. Sioutos, M.S. Tuttle, L.W. Wright et al., NCI Thesaurus: using science-based terminology to integrate cancer research results., in: *Medinfo*, 2004, pp. 33–37, doi:10.3233/978-1-60750-949-3-33.

[15] N. Sioutos, S. de Coronado, M.W. Haber, F.W. Hartel, W.-L. Shaiu and L.W. Wright, NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information, *Journal of biomedical informatics* **40**(1) (2007), 30–43, 10.1016/j.jbi.2006.02.013.

[16] G.-Q. Zhang, L. Cui, J. Teagno, D. Kaebler, S. Koroukian and R. Xu, Merging ontology navigation with query construction for web-based Medicare data exploration, *AMIA Summits on Translational Science Proceedings* **2013** (2013), 285.

[17] D.S. Scott, Domains for denotational semantics, in: *International Colloquium on Automata, Languages, and Programming*, Springer, 1982, pp. 577–610, doi:10.1007/BFb0012801.

[18] G.-Q. Zhang, *Logic of domains*, Springer Science & Business Media, 2012.

[19] S. Abramsky and A. Jung, Domain theory (1994).

[20] G. Winskel, *The formal semantics of programming languages: an introduction*, MIT press, 1993.

[21] G.-Q. Zhang, Disjunctive systems and L-domains, in: *International Colloquium on Automata, Languages, and Programming*, Springer, 1992, pp. 284–295, doi:10.1007/3-540-55719-9_81.

[22] N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D.L. Rubin, M.-A. Storey, C.G. Chute et al., BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic acids research* **37**(suppl_2) (2009), W170–W173, doi:10.1093/nar/gkp440.

[23] O. Bodenreider, Biomedical ontologies in action: role in knowledge management, data integration and decision support, *Yearbook of medical informatics* **17**(01) (2008), 67–79, doi:10.1055/s-0038-1638585.

[24] Data Analytics with SNOMED CT, https://confluence.ihtsdotools.org/display/DOCANLYT/Data+Analytics+with+SNOMED+CT, Accessed 5 May 2019.

[25] SNOMED CT Editorial Guide, https://confluence.ihtsdotools.org/display/DOCEG/SNOMED+CT+Editorial+Guide, Accessed 5 May 2019.

[26] L. Cui, S. Tao and G.-Q. Zhang, Biomedical ontology quality assurance using a big data approach, *ACM Transactions on Knowledge Discovery from Data (TKDD)* **10**(4) (2016), 41, doi:10.1145/2768830.

[27] M. Amith, Z. He, J. Bian, J.A. Lossio-Ventura and C. Tao, Assessing the practice of biomedical ontology evaluation: Gaps and opportunities, *Journal of biomedical informatics* **80** (2018), 1–13, 10.1016/j.jbi.2018.02.010.

[28] X. Zhu, J.-W. Fan, D.M. Baorto, C. Weng and J.J. Cimino, A review of auditing methods applied to the content of controlled biomedical terminologies, *Journal of biomedical informatics* **42**(3) (2009), 413–425, doi:10.1016/j.jbi.2009.03.003.

[29] G.-Q. Zhang and O. Bodenreider, Large-scale, exhaustive lattice-based structural auditing of SNOMED CT, in: *AMIA Annual Symposium Proceedings*, Vol. 2010, American Medical Informatics Association, 2010, pp. 922–926.

[30] G.-Q. Zhang and O. Bodenreider, Using SPARQL to Test for Lattices: application to quality assurance in biomedical ontologies, in: *International Semantic Web Conference*, Springer, 2010, pp. 273–288.

[31] L. Cui, W. Zhu, S. Tao, J.T. Case, O. Bodenreider and G.-Q. Zhang, Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT, *Journal of the American Medical Informatics Association* **24**(4) (2017), 788–798, doi:10.1093/jamia/ocw175.

[32] R. Abeysinghe, M.A. Brooks, J. Talbert and C. Licong, Quality assurance of NCI Thesaurus by mining structural-lexical patterns, in: *AMIA Annual Symposium Proceedings*, Vol. 2017, American Medical Informatics Association, 2017, pp. 364–373.

[33] L. Cui, O. Bodenreider, J. Shi and G.-Q. Zhang, Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs, *Journal of biomedical informatics* **78** (2018), 177–184, doi:10.1016/j.jbi.2017.12.010.

[34] G.-Q. Zhang, G. Xing and L. Cui, An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies, *Journal of biomedical informatics* **80** (2018), 106–119, 10.1016/j.jbi.2018.03.004.

[35] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill and I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), *Journal of the American Medical Informatics Association* **17**(2) (2010), 124–130, doi:10.1136/jamia.2009.000893.

[36] B. Haarbrandt, E. Tute and M. Marschollek, Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository, *Journal of biomedical informatics* **63** (2016), 277–294, doi:10.1016/j.jbi.2016.08.007.

[37] R.L. Fleurence, L.H. Curtis, R.M. Califf, R. Platt, J.V. Selby and J.S. Brown, Launching PCORnet, a national patient-centered clinical research network, *Journal of the American Medical Informatics Association* **21**(4) (2014), 578–582, doi:10.1136/amiajnl-2014-002747.

[38] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek et al., Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers, *Studies in health technology and informatics* **216** (2015), 574–578.

[39] D.A. Dean, A.L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S.S. Sahoo, C.P. Jayapandian, L. Cui, M.G. Morrical, G.-Q. Zhang and S. Redline, Scaling

up scientific discovery in sleep medicine: the National Sleep Research Resource, *Sleep* **39**(5) (2016), 1151–1164.

[40] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley and S. Redline, The National Sleep Research Resource: towards a sleep data commons, *Journal of the American Medical Informatics Association* **25**(10) (2018), 1351–1358, doi:10.1093/jamia/ocy064.

[41] S. Arabandi, C. Ogbuji, S. Redline, R. Chervin, J. Boero, R. Benca and G.-Q. Zhang, Developing a sleep domain ontology, *AMIA Clinical Research Informatics Summit* (2010), 12–13.

[42] Center for Sudden Unexpected Death in Epilepsy Research, http://sudepresearch.org, Accessed 5 May 2019.

[43] S.S. Sahoo, S.D. Lhatoo, D.K. Gupta, L. Cui, M. Zhao, C. Jayapandian, A. Bozorgi and G.-Q. Zhang, Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care, *Journal of the American Medical Informatics Association* **21**(1) (2013), 82–89, doi:10.1136/amiajnl-2013-001696.

[44] Surveillance Epidemiology and End Results (SEER) program, https://seer.cancer.gov/, Accessed 5 May 2019.

[45] NAACCR data dictionary, http://datadictionary.naaccr.org/?c=10, Accessed 5 May 2019.

[46] K.G. Pope and S. Krasner, A cookbook for using the model-view-controller user interface paradigm in Smalltalk-80, *Journal of Object-Oriented Programming* **1** (1988).

[47] S. Tao, B.L. Walter, S. Gu and G.-Q. Zhang, Web-Interface-Driven Development for Neuro3D, a Clinical Data Capture and Decision Support System for Deep Brain Stimulation, in: *International Conference on Health Information Science*, Springer, 2016, pp. 31–42, doi:10.1007/978-3-319-48335-1_4.

[48] L. Cui, N. Zeng, M. Kim, R. Mueller, E.R. Hankosky, S. Redline and G.-Q. Zhang, X-search: an open access interface for cross-cohort exploration of the National Sleep Research Resource, *BMC medical informatics and decision making* **18**(1) (2018), 99, doi:10.1186/s12911-018-0682-y.

[49] G.-Q. Zhang, L. Cui, S. Lhatoo, S.U. Schuele and S.S. Sahoo, MEDCIS: multi-modality epilepsy data capture and integration system, in: *AMIA Annual Symposium Proceedings*, Vol. 2014, American Medical Informatics Association, 2014, p. 1248.

[50] S. Tao, L. Cui, X. Wu and G.-Q. Zhang, Facilitating Cohort Discovery by Enhancing Ontology Exploration, Query Management and Query Sharing for Large Clinical Data Repositories, in: *AMIA Annual Symposium Proceedings*, Vol. 2017, American Medical Informatics Association, 2017, p. 1685.