Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 17(4): 4098–4114.
DOI: 10.3934/mbe.2020227
Received: 31 December 2019
Accepted: 18 May 2020
Published: 08 June 2020

Research article

PMO: A knowledge representation model towards precision medicine

Li Hou†, Meng Wu†, Hongyu Kang, Si Zheng, Liu Shen, Qing Qian and Jiao Li*

Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100020, China

† These authors contributed to this work equally.

* **Correspondence:** Email: li.jiao@imicams.ac.cn.

Abstract: With the rapid development of biomedical technology, amounts of data in the field of precision medicine (PM) are growing exponentially. Valuable knowledge is included in scattered data in which meaningful biomedical entities and their semantic relationships are buried. Therefore, it is necessary to develop a knowledge representation model like ontology to formally represent the relationships among diseases, phenotypes, genes, mutations, drugs, etc. and achieve effective integration of heterogeneous data. On basis of existing work, our study focus on solving the following issues: (i) Selecting the primary entities in PM domain; (ii) collecting and integrating biomedical vocabularies related to the above entities; (iii) defining and normalizing semantic relationships among these entities. We proposed a semi-automated method which improved the original Ontology Development 101 method to build the Precision Medicine Ontology (PMO), including defining the scope of the PMO according to the definition of PM, collecting terms from different biomedical resources, integrating and normalizing the terms by a combination of machine and manual work, defining the annotation properties, reusing existing ontologies and taxonomies, defining semantic relationships, evaluating PMO and creating the PMO website. Finally, the Precision Medicine Vocabulary (PMV) contains 4.53 million terms collected from 62 biomedical vocabularies, and the PMO includes eleven branches of PM concepts such as disease, chemical and drug, phenotype, gene, mutation, gene product and cell, described by 93 semantic relationships among them. PMO is an open, extensible ontology of PM, all of the terms and relationships in which could be obtained from the PMO website (http://www.phoc.org.cn/pmo/). Compared to existing project, our work has brought a broader and deeper coverage of mutation, gene and gene product, which enriches the semantic type and vocabulary in PM domain and benefits all users in terms of medical literature annotation, text mining and knowledge base construction.

Keywords: biomedical ontology; precision medicine; semantic web; controlled vocabulary; taxonomy

1. Introduction

Precision medicine (PM) is an approach to disease treatment and prevention that seeks to maximize effectiveness by taking into account individual variability in genes, environment, and lifestyle [1]. A PM ecosystem also links patients, providers, clinical laboratories and researchers [2]. With the rapid development of biomedical technology, amounts of heterogeneous clinical data and scientific information are growing exponentially, in which valuable knowledge like meaningful biomedical entities and their semantic relationships are buried. How to make sense of the knowledge and use them to support PM is a crucial issue. Through integrating, classifying and standardizing these biomedical entities from different resources and specifying relationships among them, terminologies, classifications and ontologies are able to provide computational formats of differing degrees of sophistication, which allow analysis of large amounts of data and precise classification of a patient [3]. Building these knowledge models are essential and necessary for the study of PM.

Terminologies has been used to integrate heterogeneous data for a long history, such as the Unified Medical Language System (UMLS), which is a large repository of biomedical vocabularies [4]. The three knowledge sources of UMLS includes the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. In the 2019AB release of the UMLS, the Metathesaurus integrates concepts, concept names, and other attributes from more than 200 electronic versions of numerous thesauri, classifications, code sets, etc. The Semantic Network provides a consistent categorization of all concepts represented in the Metathesaurus and a set of useful relationships. It comprises 133 semantic types of organisms, anatomical structures, biological functions, chemicals, events and physical objects, with 54 semantic relationships such as contains, co-occurs, affects, diagnoses, adjacent to, and so on. A powerful use of the UMLS is linking health information, medical terms, drug names, and billing codes across different computer systems. Other uses include search engine retrieval, data mining, public health statistics reporting, and terminology research [4]. But there is a lack of integration of large data at the level of disease's molecular mechanisms such as gene, mutation, gene product, etc., the semantic relationships in UMLS also lack the detailed relationships between these entities [5], which constitute an important aspect of PM. Further integration and standardization for the concepts of the molecular biology and the more detailed associations between them are especially necessary in the study of PM.

Ontologies differ from terminologies in that ontologies define relationships between concepts in a way that allows computational logical reasoning, enabling the drawing of conclusions from related assertions [6]. An ontology is a formal explicit description of concepts in a domain, the properties of each concept describing its various features and attributes, and restrictions on properties [7]. An ontology can provide a vocabulary, standard identifiers, metadata, and machine-readable axioms and definitions for classes and relationships that represent the phenomena within a domain. Up to now, ontologies are widely used in biological and biomedical research with the advantage of facilitating data integration, access and analysis [8]. Researchers use ontologies to annotate data with ontology terms, enabling improved data integration and interoperability across disparate datasets [9]. For PM, an ontology can provide: (1) Data support for precision medical text mining and knowledge graph construction; (2) technical support for the management of PM knowledge base (3) application support for scientific research and clinical practice in PM. Therefore, we think an ontology will be a suitable model to represent the knowledge of PM.

Multiple ontologies for various domains of biomedicine have been built, such as Gene Ontology (GO), a comprehensive resource of computable knowledge regarding the functions of genes and gene products [10] and Disease Ontology (DO) [11], an ontology for human diseases with human disease terms, phenotypic characteristics and related medical vocabulary disease concepts. The Human Phenotype Ontology (HPO) provides comprehensive bioinformatics resources for the analysis of

human diseases and phenotypes, offering a computational bridge between genome biology and clinical medicine [12]. In addition to these universal medical domain ontologies, researchers have created more targeted domain ontologies according to individual work requirements. The Drug Target Ontology [13] was developed to integrate and analyze drug discovery data of various resources based on classifications and annotations of drug protein targets, including related information among proteins, genes, protein domains, binding sites, small molecule drugs, mechanisms of action and many other types of information. The Non-Coding RNA Ontology (NCRO) [14] is a comprehensive resource for the unification of non-coding RNA biology. These ontologies cover portions of PM domain in some degree. A more comprehensive ontology for representing the data types in various aspects of clinical, pathological, and molecular studies and the relationships between them is significant and convenient to the PM study.

For the PM, the combination of massive data and affordable high-capacity computing of ontology provides an opportunity for unprecedented discovery of association and, increasingly, causal reasoning to gain diagnostic and therapeutic insight [3]. Some works have been done, like the Precision Medicine Ontology (PreMedOnto) [15], which consists of 543 annotated classes and 10 properties. PreMedOnto reuses terms and concepts from other ontologies and maps the terms extracted from domain specific texts collected from PubMed repository to the existing concepts. The PreMedOnto is able to capture and represent the semantics of the PM domain with high precision and significance, but in which the classification is not comprehensive and the relationships are not every abundant. For example, there is only one class "Protein" under the class "Gene Product" [16]. Ontology of Precision Medicine and Investigation (OPMI) [17] is an application ontology to support PM and related investigations, which also reuses, aligns, and integrates related terms from existing ontologies in the OBO ontology library. The Basic Formal Ontology (BFO), which is a genuine upper ontology, is reused for organizing the top-level classes in OPMI. The OPMI only contains the data in the study of kidney disease, but the representation model of the entities associated with PM is a significant reference for the ontology construction in PM domain. Currently, there is still a lack of large ontology in PM domain which contains comprehensive classification and relationships and covers various kinds of disease simultaneously.

In this work, the Precision Medicine Ontology (PMO) was developed as a standard ontology for integrating and representing the data in the human PM domain with consistent, reusable and sustainable descriptions of human diseases, genomic and molecular features and phenotypic characteristics through collaborative efforts from multidisciplinary researchers. The PMO contains eleven branches of PM concepts such as disease, chemical and drug, phenotype, gene, mutation, gene product and cell linked by 93 semantic relationships among them. Meanwhile, we built a Precision Medicine Vocabulary (PMV) containing 4.53 million terms collected from 62 biomedical vocabularies. The building process indicated the scalability and flexibility of the representation model we designed. All of the terms and relationships in PMO could be obtained from the website (http://www.phoc.org.cn/pmo/).

2. Methods

The commonly used construction techniques for ontologies mainly include the manual method, reusing the existing ontologies and the automatic method. Many study have been published on manual construction methods, including the ontology development 101 [7], skeleton [18], SENSUS [19], KACTUS [20], IDEF5 [21], METHODOLOGY [22] and TOVE [23] methods. Manual construction tools for ontology such as protégé[24], generally support editing, visualization, reasoning, refactoring

of ontology, by which users can complete ontology construction through a series of manual operations. The approaches and tools for constructing ontology manually are not applicable to the ontologies with massive data. Reusing the existing ontologies is a general operation in biomedical domain, as there are many mature biomedical ontologies that have been built as mentioned above. The automatic methods (also known as ontology learning) still face many challenges. Most of the techniques and tools used in state-of-the-art ontology learning methodologies are designed for smaller data sets. The quality of learned ontologies is also affected by the human intervention [25].

In this paper, we proposed a semi-automated method which improved the original Ontology Development 101 method by utilizing the automatic work for large-scale data extraction and integration. We chose the Ontology Development 101 method because it provided a very concrete and applicable guidance in each step and its iterative design allowed the ontology developers revise an ontology easily [26]. The manual work mainly focuses on data analyzing, ontology designing and semantic relationships defining. Our method for PMO construction followed the steps from scope definition to website deployment as following:

Step 1. Defining the scope of the PMO according to the definition of PM.
Step 2. Extracting and collecting terms from different ontologies, vocabularies and databases.
Step 3. Integrating and normalizing the heterogeneous biomedical resources.
Step 4. Defining the annotation properties of PMO classes.
Step 5. Reusing the hierarchical structure of existing ontologies and taxonomies.
Step 6. Defining semantic relationships by manual method.
Step 7. Evaluating PMO
Step 8. Creating the PMO website.

2.1 Defining the scope of the PMO

The PMO is an open source ontological knowledge representation model about the data in the field of PM. The concept of PM was firstly profiled by a publication of the National Research Council, which states a new data network that integrates emerging research on the molecular makeup of diseases with clinical data for individual patients could drive the development of a more accurate classification of diseases and ultimately enhance diagnosis and treatment [27]. Meanwhile, the PM research initiative in America, which is called ALL of US research program [28], focuses on the intersection of environment, lifestyle, and biology. To provide a better description of data in the PM field, signs and symptoms, the genome and other data related to the disease's molecular mechanism and etiology as well as the relationships among them were considered in this study.

The scope was also defined and extended to meet the needs of the whole project of precision medicine knowledge base. The Precision Medicine Knowledge Base, the Chinese National Key Research and Development Program, aims to construct a reliable knowledge base of PM for massive data analysis and integration. The PMO supports the identification and semantic integration of genes, diseases, drugs, mutations, phenotypes, pathways, and so on, which represent the key scope of data in the PM knowledge base.

The PMO was organized into eleven top classes to represent anatomical structure, gene, gene product, mutation, cell, disease, phenotypic abnormality, gene function, biomedical pathway, biological function, chemical, and drug with uniform identification. The PMO project continues to improve the representation of all data in the human PM domain, with the addition of new PM terms as needed for curation, term requests and collaborative development. Crucial efforts are underway to strengthen and expand the PMO's representation of semantic relationships among these classes to

describe disease pathogenesis in a more precise way. In the future, the scope of the PMO will be expanded by taking into account data on the exposome, samples and clinical trials, which could improve outcome translation into clinical practice.

2.2 Collecting and organizing the vocabularies for the PMO

A main feature that the ontology provide is a set of terms associated with the classes and relationships, which are usually referred to as labels. In ontologies, labels may be provided in multiple languages, and a primary label may be distinguished from secondary labels or synonyms for a given class or relationship [8]. We collected vocabularies from both comprehensive and specialized vocabularies in the PM field to build the PMV for the PMO (as the controlled vocabulary). Then, we organized the vocabularies into the standard concept-term structure for managing the vocabularies and linking them with ontology.

We collected vocabularies from the UMLS, since some semantic types in the UMLS were considered necessary in PMO scope and it has supplied a comprehensive integration of many biomedical vocabularies. First, we extracted the UMLS concepts whose semantic types were within the scope, including Anatomical Structure, Chemical, Clinical Drug, and Sign or Symptom and so on. According to the current need of the PMO, we removed the concepts of non-human species by mapping the terms in UMLS with the terms of other species under the Eukaryota (Tree Number: B01) in MeSH tree, and removed the non-English languages terms by utilizing the language of term (LAT) in UMLS. Finally, after removing the source vocabularies in which terms are very few and whose subjects are not related to our scope, 56 source vocabularies remained. The foundational vocabulary contained both comprehensive vocabularies such as Medical Subject Headings (MeSH), National Cancer Institute Thesaurus (NCIt) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and databases in specific domains such as HUGO Gene Nomenclature Committee (HGNC) and Online Mendelian Inheritance in Man (OMIM) for gene, Human Phenotype Ontology (HPO) for human phenotype, DrugBank and RxNorm for drug.

We followed and simplified the assignment method of several unique identifiers for the concepts in the UMLS to reorganize the selected data. This mechanism is also utilized for unique identifier setting and term management in the whole vocabulary system. In this mechanism, terms from different sources with same semantic meaning were merged into one concept, and only one term among them was set as the preferred term for the concept. The metadata of the vocabulary included name, IDs, language, preferred term tag, Term Type in Source, the source's abbreviation and source code. Term Type in Source (TTY), which is the term type used in source description in the UMLS, was reused by matching the terms to appropriate term types and extended by defining new term types to record the features of the original data. In addition to the main concept table, there were other tables describing the source and version of each vocabulary and semantic type for every concept in the PMV.

Due to the sources of the concepts and terms are not only UMLS, a set of new ID rules were created for PMV. MCID was used for concept identification, MAID for term identification, PMOID for class identification, and RID for relationship identification. So each MCID may be linked to at least one MAID. We also defined the rules for coding the unique identifier of each term/concept, e.g. the Term ID was called "MAID" and its value would be expressed as "MA00019781". The digits of each code were designed according to the actual demand of the knowledge base (Table 1).

Table 1. Vocabulary coding formats.

Name	Abbreviation	Length	Format	Example
Concept ID	MCID	10 digits	MC + 8 digits	MC00001175
Term ID	MAID		MA + 8 digits	MA00019781
Class ID	PMOID	12 digits	PMO: + 8 digits	PMO:00000035
Relationship ID	RID	9 digits	R + 8 digits	R00000001

2.3. Integrating the heterogeneous biomedical resources

To provide a better representation of the data in PM domain, we integrated biomedical resources such as DrugBank, ClinVar, National Center for Biotechnology Information (NCBI) Gene, Disease Ontology, Human Phenotype Ontology with the foundational vocabulary by utilizing a series of mapping and integration strategies (Figure 1). We chose these databases because in which the data types are covered by the scope defined above but the data are fewer in UMLS. We extracted the gene data from NCBI gene database which provides gene sequence information of multiple species, including sequence, expression, structure, function and reference, and the unique identification of gene Entrez_ID is commonly used in all databases developed by NCBI [29]. ClinVar is a public database of human genetic variants associated with disease, in which the variants are curated by experts to provide a more accurate information about the relationships between genotypes and phenotypes [30]. Disease Ontology has been proved to be resource rich in cross-references with other disease vocabularies. DO terms and IDs are also wildly used in many algorithms, computational tools and biomedical resources [11]. Among them, ClinVar, NCBI Gene, Disease Ontology are not involved in UMLS. DrugBank and HPO have been included in UMLS already, but we found that UMLS integrates them incompletely. Further mapping and integration of them are necessary in this work.

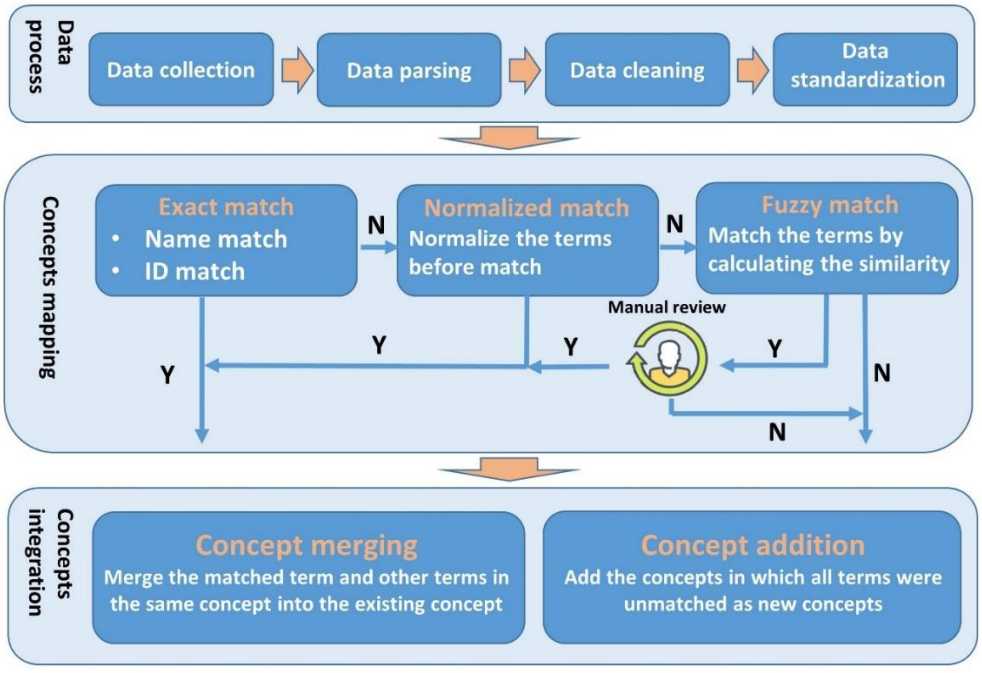

Figure 1. The detailed integration workflow of the heterogeneous biomedical resources. "Y" and "N" respectively represent the positive and negative results of the judgement.

Each resource was first parsed and transformed into a concept-term form according to its content and structure. In order to match terms in new resource with existing terms in PMV as many as possible, three match strategies were employed in turn. The exact match utilized the term name, ID and resource name to provide an accurate mapping. The Norm tool [31], which is one of the lexical tools of the UMLS, was used in the normalized match process. The tool FuzzyWuzzy was used in fuzzy match process, in which the Levenshtein Distance method was used to calculate the similarity of the words. After the fuzzy match, manual review process was performed. Biomedical experts checked the top three matching results of the fuzzy match and picked the most correct one or failed them all. The matched terms and other terms in the same concept were merged into the existing concept in the PMV. The concepts in which all terms were unmatched were added into the PMV as new concepts. For instance, the gene "A1BG" (NCBI Gene ID:1) has five synonyms, "A1B", "ABG", "HYST2477", "alpha-1-B glycoprotein" and "GAB" in NCBI Gene. We represented this gene data in a format of a concept with six terms. The term "A1BG" in this concept was matched to the term "A1BG" in the concept "MC00493168" in PMV after the exact match process. Then, these six terms were all merged into the concept "MC00688619" as new terms.

In the concept match process of DO, 9904 disease concepts were mapped to the existing disease concepts in PMV at the exact match step in which 8004 were mapped through reference ID match and 1900 were mapped through exact string match. Then, 12 disease concepts were mapped by utilizing the norm tool. After that, 141 were mapped by utilizing the fuzzy match tool and 51 were approved after the manual review. Meanwhile, 4571 drug concepts and 150,629 terms in DrugBank were added into the PMV, which had not previously been identified and integrated in the UMLS. NCBI Gene and ClinVar, which are new resources, were added into the PMV through mapping with existing genes and mutations mainly in OMIM and HUGO. It provided additional 21,172 concepts and 220,328 terms in Gene and 294,712 concepts and 316,630 terms in Mutation.

2.4. Defining the annotation properties of classes

The annotation properties were defined to describe the PMO in many dimensions. PMOID is the unique identifier of PMO classes. MCID links the PMO and PMV by mapping concepts in the PMV with the classes in the PMO. Definition provides textual definitions of classes and relationships, most of which were inherited from well-known biomedical databases. Database_cross_reference provides the IDs mapped with 62 biomedical databases. The diversity of cross-reference sources indicates the interoperability of the PMO. Meanwhile, the PMO integrates and connects synonyms of a class from different databases based on the terms of the concept in PMV. The primary annotation properties and definitions are listed in Table 2 below.

2.5. Reusing the hierarchical structure of existing ontologies and taxonomies

As the scope of the PMO is broad and diverse, some classification systems and terms in existing and well-known biomedical resources were reused as needed. According to the eleven top classes considered above, we selected and reused the terms and hierarchies from well-known biomedical resources such as MeSH, NCIt, UMLS, HPO, Variation Ontology (VariO), and NCBI Gene. These terms and hierarchies were jointly integrated into a standard ontology structure (Figure 2).

Table 2. The annotation properties of PMO.

Annotation property	Definition
PMOID	The unique identifier of the class in PMO
MCID	The identifiers of the corresponding concepts of the class in PMV
MRID	The unique identifier of the relationship in PMO
Name	The common name of the resource in PMO
Tree Number	The hierarchy of the resource in PMO tree
Definition	The definition of the resource given by experts of PMO or obtained from other biomedical database
Database_Cross_Reference	The IDs mapped with the resource in other databases
Synonym	The names of the resource in other databases
Subclass_of	The superclasses of the resource
Example	The example of the relationship appearing in biomedical text
Source of Example	The source of the example, usually PubMed

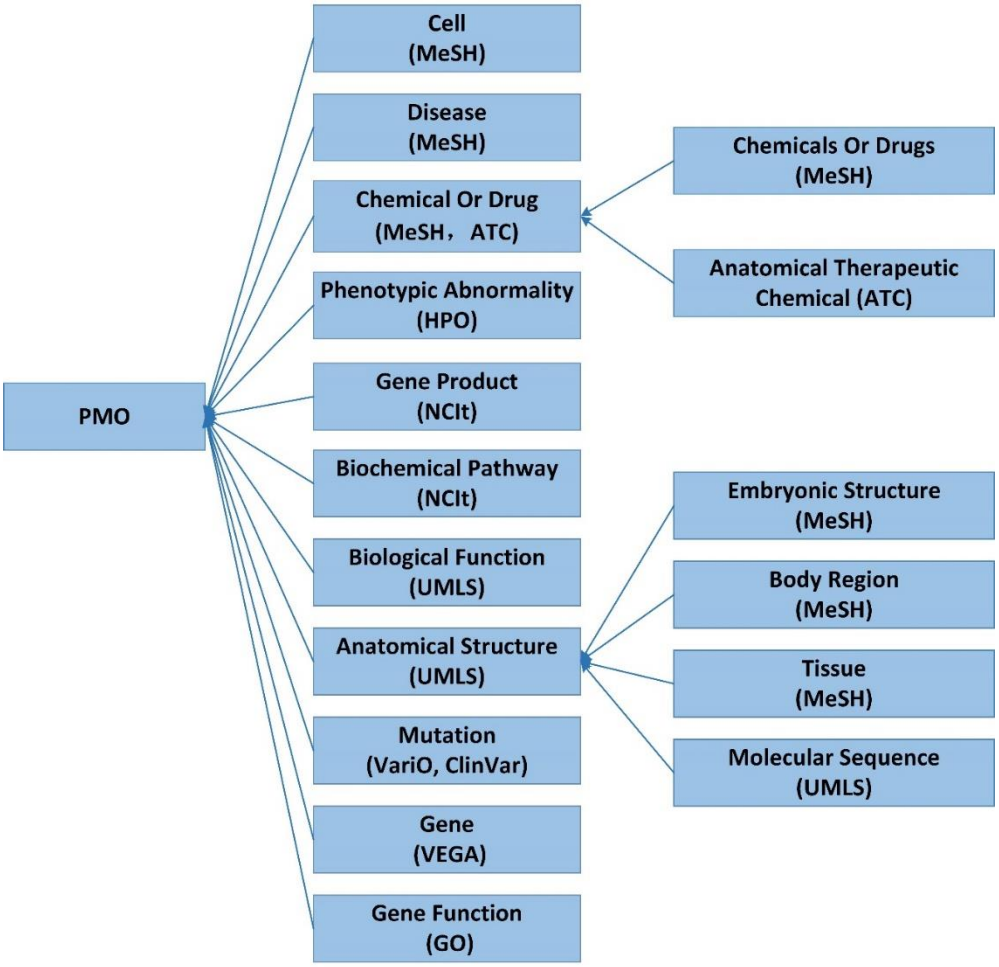

Figure 2. The top level PMO hierarchical structure and key ontology terms. Terms and hierarchies reused from other biomedical resources are indicated by the abbreviations inside the parentheses. All the arrows indicate the "Subclass_of" relationship.

From top to bottom, considering that MeSH which provides an effective classification for medical term is the preferred source of the UMLS, so the hierarchies under the top classes such as cell, disease and chemical or drug were constructed by referring to the MeSH. A broader class relationship was defined in PMO to link the upper class and lower class between which the relationship is less rigorous Subclass_of relationship in MeSH. The hierarchies under gene product and biomedical pathway were adopted from NCIt, which has a good coverage of data on molecular mechanisms in the development of cancer. In the branch of phenotypic abnormality, the preferred name and subclasses of the class "Phenotypic Abnormality" were adopted from HPO. The terms and hierarchies under the gene function branch were reused from Gene Ontology, in which terms and IDs are widely used in many algorithms, computational tools and biomedical resources [32]. Others, such as hierarchies under biologic function and anatomical structure were directly adopted from the UMLS to supply an adaptive classification system for PMV vocabularies.

From bottom to up, mutation terms were organized based on the data types in the source database of ClinVar. The data types were mapped to the classes in VariO, which were reused in the hierarchy of the mutation branch. The terms of gene were organized based on the gene classification of vertebrate genome annotation (VEGA) database. Specific genes and mutations were appended on the PMO as instances of the corresponding class. For example, in the NCBI Gene database, genes are classified into several types, including protein-coding, ribosomal RNA (rRNA), transfer RNA (tRNA) and so on. The PMO inherited these data types as classes for organizing gene data better.

It's worth mentioning that the notion of Drug refers to chemical substances in PMO, the brand names of drugs were supplied as the synonyms. We used the Lexical alignment method [33] to process an elementary class mapping. The identical classes from different biomedical resources were linked through the owl: equivalentClass property if they were mapped to the same concept in PMV, such as the same terms of drugs in ATC and MeSH.

In summary, all the efforts made above were to map classes in the PMO with terms in the PMV as many as possible, so that a fine classification system of the PMV which supplies abundant synonyms and cross-references can be achieved. When we integrated different classifications, some contradictions occurred. In the contradiction of hierarchy, class A is the subclass of Class B in a classification, class B is the subclass of Class A in another classification. In this case, we would invite the domain experts to review the contradiction manually, and decided which hierarchy would be used in PMO.

2.6. Defining semantic relationships in the PMO

Defining and normalizing semantic relationships among biomedical entities inherited from different resources is an important part of the PMO. Due to the complexity of relationships among concepts, as an early step, we focused on extracting and defining relationships among top classes by utilizing domains and ranges in the ontology. To represent the relationship in PM domain, we preferred the more detailed relationships between the entities of molecular biology. Currently, we have manually curated and summarized 93 semantic relationships for 7 PMO top classes (Table 3) by referring to the relationships curated from biomedical literature by the experts from QIAGEN Ingenuity Pathway Analysis (QIAGEN IPA) Systems [34,35]. Furthermore, we designed the hierarchy of the relationships to organize them, and provided a standard name and a definition for each relationship. The hierarchical relationships were transformed into a standard OWL format to build a special ontology of relationships for enabling a unique identification and extensible representation model for relationships in the PM field. Each relationship description has both Chinese and English versions.

Table 3. Selected Object Properties of Gene domain in the PMO.

RID	Relationship _name	Domain	Range	Definition	Hierarchy
R00000001	is biomarker of	Gene	Disease	A gene influences or predicts the incidence of outcome or disease.	First
R00000002	is biomarker-efficacy of	Gene	Disease	A gene can be used to measure the efficacy of drugs or therapeutic methods in the treatment of a disease.	Second
R00000003	is biomarker-diagnosis of	Gene	Disease	A gene can be used to diagnose a disease.	Second
R00000004	is biomarker-prognosis of	Gene	Disease	A gene can be used to predict the probable course or outcome of a disease.	Second
R00000005	is biomarker-response to therapy of	Gene	Disease	A gene can be used for measuring the response of a disease to particular therapy.	Second
R00000006	is biomarker-unspecified application of	Gene	Disease	A gene is a potential biomarker of a disease: the gene (when mutated or aberrantly expressed) is associated with a disease, but the precise function is unclear.	Second

Note: RID: the unique identifier of the semantic relationship; Domain: the subject of the semantic relationship; Range: the object of the semantic relationship.

2.7. Evaluating PMO

The evaluation of an ontology generally comprises assessing its inner features and utility. The inner features include the clarity, accuracy and consistency of the ontology. In terms of clarity, all PMO terms were given non-ambiguous labels for a clear and effective interaction with users and terms from other databases. No abbreviations or messy codes were used. For accuracy, the PMO reused the latest and authoritative databases to the greatest extent to supply the most recognized representation model for PM. Meanwhile, many domain experts provided guidance during PMO construction in top classes setting, resource integration, and other steps. In consistency check, we defined rules and used the tool ROBOT [36] to identify contradictions and redundancy in class names, hierarchies and properties to prevent the inconsistency errors, such as circulatory errors, inheritance relationship errors and hierarchy errors. Through the evaluation of PMO, we perfected the ontology constantly in an iterative mode.

The PMO was constructed by following the OBO Foundry principles, which are also a series of standards for ontology evaluation [37]. Though the alignment, we have meet the criteria by providing openness, common format, Uniform Resource Identifier (URI)/identifier space, versioning, clear scope, relationships, collaboration, contact person, naming conventions and maintenance. Besides, we have supplied some textual definitions extracted from the source databases, even more definitions for classes have to be added. And the documentation of the PMO will be covered in our future work.

2.8. Creating the PMO website and updating

We used Web 2.0 and semantic web technologies to display all the terms and relationships of PMO. With the help of these technologies, the detailed information of terms in the whole ontology tree can be presented well. PMO term metadata and visualization are on a single page with multiple tabs. The metadata refer to the object properties and annotation properties in PMO in the website. It allows the pages of metadata browsing, search results or visualizations to remain while the ontology is further explored (Figure 3).

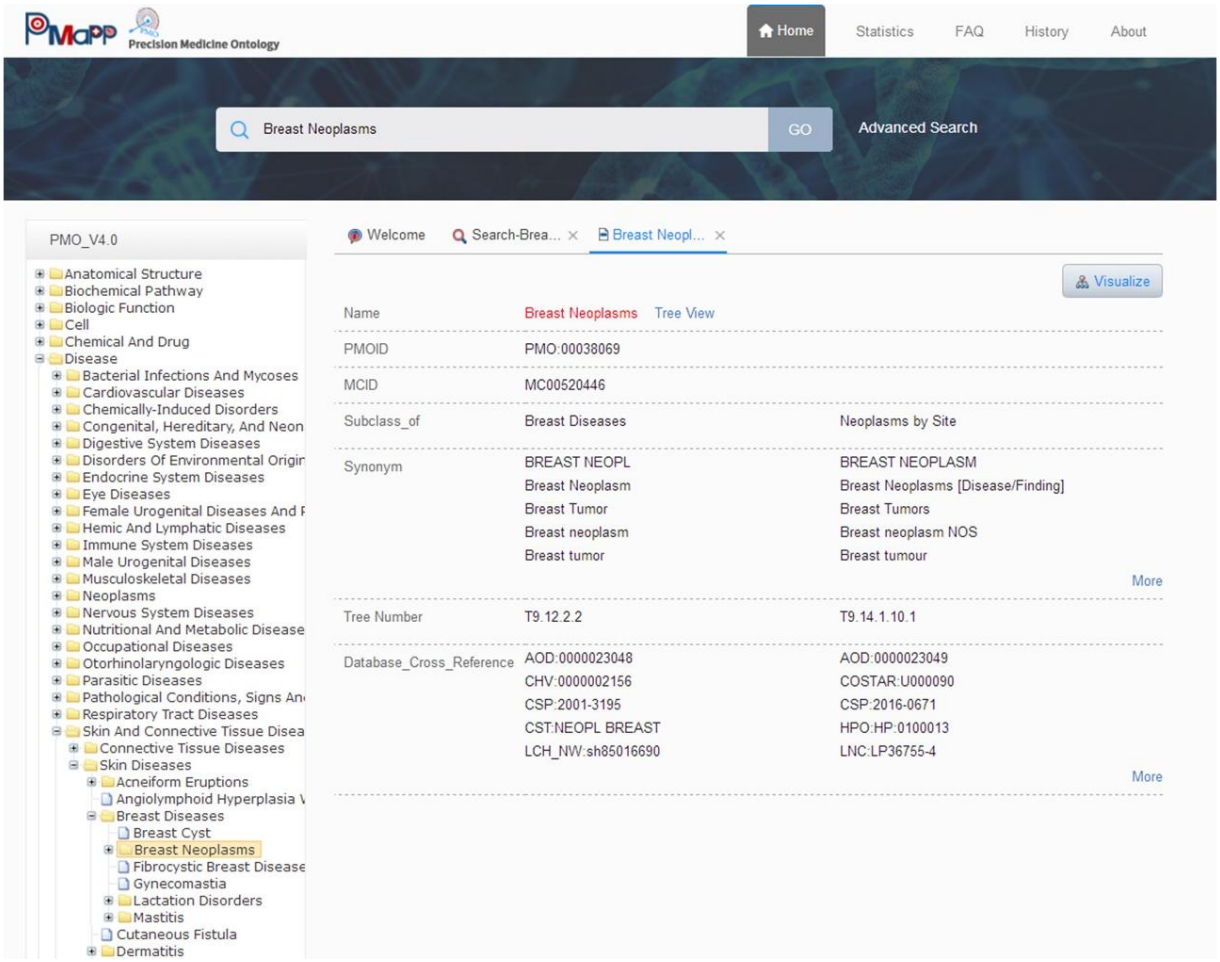

Figure 3. The PMO website. The properties for PMO term of Breast Neoplasms are displayed.

The layout of the PMO homepage can be subdivided into three distinct sections. The "Search Panel" provides basic and advanced queries on the PMO. Basic Search provides searches against the two fields of Name and Synonym and Advanced Search allows the user to generate complex Boolean queries on selectable fields of the ontology. The "Navigation Panel" provides a hierarchical structure view of the PMO. The subclasses can be expanded through a single-click on the indicated class, and detailed content of the class can be obtained through a click on the class. The "Content panel" contains the metadata of the class, search results, tree view and visualization of the class organized through multiple tabs. The visualization function provides an intuitive form for representing classes and the relationships among the subclasses and superclasses of the selected class by utilizing connective nodes and edges in different colors.

In the aspect of versioning, intensive effort has been made to design an updating mechanism to keep the PMO resource sustainable and combine the updates of PMV and PMO together. The PMV and PMO are updated annually for tracking the evolution of the constituent resources. For PMV, we update the concepts and preferred terms by identified terms changing types and preferred terms changing modes of concepts in the source vocabulary, which is the method mentioned in our previous work [38]. For PMO, the hierarchical structure is modified according to the updates of the reused ontologies and taxonomies. Once PMV or PMO changes will update the PMO because of the close connection between them. When PMV term changes, the PMO content will be updated as the values of Synonym and Database_Cross_Reference are from the PMV. When the structure of the PMO changes, it needs to map with the PMV again to obtain the values of corresponding properties according to the new class names and hierarchies. Up to now, four main versions have been released upon PMO, whose details can be viewed on the PMO website.

3. Results and discussions

3.1. Statistics and analysis

As mentioned above, we built a controlled vocabulary for integrating and standardizing the terms used in PM research. The latest version of the PMV contains 2,609,748 concepts and 4,636,459 terms extracted from 62 biomedical vocabularies, which covers eleven top semantic types in PM domain. The large number of source vocabularies included in the PMV indicates its role in interoperability in the PM field. To assess the novel knowledge representation model in an effective and comprehensive way, we compared the concepts in the PMV and those in the UMLS in dimension of quantity (Table 4).

Table 4. Numbers of concepts in the PMV and UMLS under eleven top semantic types.

TOP CLASS	Numbers in PMV	Numbers in UMLS
Anatomical Structure	99,587	197,162
Phenotypic Abnormality	39,505	13,854
Biochemical Pathway	1204	3664
Cell	5462	5,570
Biologic Function	224,111	246,575
Chemical and Drug	614,498	975,604
Disease	146,840	141,314
Gene	82,289	46,948
Mutation	320,753	25,715
Gene Product	360,117	143,375
Gene Function	715,382	65,092

By comparing the statistics in eleven different semantic types, a number of observations appeared. The UMLS has a comprehensive vocabulary, which includes multiple species and languages. Therefore, it is not surprising that the UMLS has more concepts in some semantic types compared with the PMV. But, the numbers of concepts of gene, mutation, gene product and gene function in PMV are more than those in UMLS. For the phenotypic abnormality, of which the scope is unclear in UMLS, because there is not a special semantic type for phenotypic abnormality in UMLS.

In PMV, we defined the scope and reorganized more concepts associated with phenotypic abnormality. Therefore, it indicates that the PMV may provide better applicability in more application scenes of PM, such as text mining and knowledge base construction. Although, the PMV possesses good coverage of concepts in PM domain, it provides only moderate coverage of concepts in basic medicine domain. It indicates the areas of strength and improvement for the PMV to cover additional concepts.

Currently PMV data are stored in MySQL database for management. The data of main concepts, semantic types, and other related information are all organized in the table formats. The vocabulary provides good scalability and flexibility for users. The PMV will be perfected continuously and opened in various data formats soon. Users can obtain all the terms in PMV, and rebuild their own vocabularies based on the PMV according to their personal needs, such as merging other vocabularies or extracting existing terms from the PMV.

The terms in PMV are organized into the PMO in the form of class or instance, statistics was performed of the number of subclasses under 11 classes (Table 5). We also followed the Open Biomedical Ontologies (OBO) Foundry guidelines [39] for sustainable and international development of the PMO. PMO are available under the Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0) in the format of OWL. The PMO file could be obtained from the website (http://www.phoc.org.cn/pmo/).

Table 5. PMO statistics based on eleven top classes.

Top eleven classes	Number of Classes	Number of Instances
Biochemical Pathway	763	–
Anatomical Structure	563	–
Mutation	16	320,753
Phenotypic Abnormality	13,896	–
Gene	51	82,289
Disease	4,678	–
Gene Product	25	360,117
Chemical and Drug	11,887	–
Cell	571	–
Biologic Function	11	–
Gene Function	45,022	–

3.2. Utility

The PMO has been used in other subprojects under the Precision Medicine Knowledge Base (PMKB) project, which is China's national key research and development program. In the subproject of construction of PM text knowledge network, PMO classes and relationships are being used in text corpus construction, named entity recognition (NER) and semantic relationship extraction. For example, mutation terms in the PMO are used as the dictionary for mutation name recognition and standardization. The PMO also provides unified term identification and integration for the subproject of construction of PM knowledge graph. In the subproject of automated annotation and manual curation of PM knowledge, the PMO is used as the vocabulary integrated in the dictionary management system. In the PMKB platform, PMO is used for providing annotations for the biomedical data submitted by user, and supporting the biological network analysis and enrichment analysis based on PMO [40].

In the above applications, the PMO has given enough support in various tasks and whose quality was improved in fulfilling these tasks simultaneously, which shows that the PMO plays a supporting role as an ontology.

4. Conclusions

In this study, we developed an innovative method integrating a large amount of terms from heterogeneous databases or resources, including preexisting ontologies and databases about disease, chemical, mutation and pathway. Automated construction and alignment with other related biomedical resources will expand the knowledge base much needed for the future biomedical informatics development. On the other hand, we defined semantic relationships in the PMO, which would play an important role in future scientific discoveries. We also developed the PMO website, which supplies the detailed display and the easier browse and query for the terms and relationships in PMO. The comparative statistics show the better coverage of PMV compared with UMLS on concepts in PM domain. Now PMO is used in text mining and knowledge base construction for identifying gene, mutation, chemical, disease and other biomedical terms which actively promotes efficient entity identification and standard data representation in the field of PM.

There are still some limitations in this work, which should be addressed in future work.

(1) As we processed a simple lexical alignment method for class mapping in reusing the hierarchical structure of existing ontologies and taxonomies of PMO, the effect is not optimal. Using the state of the art method of ontology mapping will be our future work.

(2) The evaluation of the usability of our work are elementary. The theoretical framework [41] will be used to evaluate the ontology in a more standard and systematic way.

(3) The relationships between specific instances are not provided in PMO, which will be supplemented through logical definitions to meet the need of the research and clinical applications and realize reasoning by being extracted from the biomedical databases and literatures.

In the future, the semantic lexicon model like Ontolex [42] or SKOS [43] could be our choice for vocabulary organization and sharing to translate PMV into a standard semantic model and provide a better connection between PMO and PMV. We will align PMO to an upper ontology such as BFO for inheriting the properties of the appropriate abstract classes in it. Meanwhile, more efforts will be made to evaluate the PMO's usability in the knowledge base and the community, to enrich more instances in the PMO and to extend its scope for covering other biomedical entities constantly. More semantic relationships and axioms will be discovered and defined to explain the pathogenic mechanism and treatment of disease to improve the computable ability of PMO. The Graph Database will be used to store the PMO and to provide several robust and fast mechanisms to retrieve individual nodes in the PMO website. As the PMO is an on-going community-driven project, it will continue to grow to overcome challenges that surface in the translational biomedical research.

Acknowledgments

We are grateful to the referees and editors for their valuable comments and suggestions for this paper. This work is supported by the National Key Research and Development Program of China (Grant No. 2016YFC0901901), the National Population and Health Scientific Data Sharing Program of China, the National Engineering Laboratory for Internet Medical System and Application (Grant No. NELIMSA2018P02), National Key Laboratory of Technology and Standards in Press and Publication Industry "Medical Fusion Publishing and Knowledge Technology Key Laboratory", the Fundamental Research Funds for the Central Universities (3332019088) and the China Knowledge Centre for

Engineering Sciences and Technology Program (CKCEST-2020-1-14).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. K. Hudson, R. Lifton, B. Patrick-Lake, E. G. Burchard, T. Coles, R. Collins, et al., The precision medicine initiative cohort program - Building a Research Foundation for 21st Century Medicine, *Precis. Med. Initiative Work. Group Rep. Advis. Comm. Dir.*, **2015** (2015).
2. G. S. Ginsburg, K. A. Phillips, Precision medicine: from science to value, *Health Aff.*, **37** (2018), 694–701.
3. M. A. Haendel, C. G. Chute, P. N. Robinson, Classification, Ontology, and Precision Medicine, *N. Engl. J. Med.*, **379** (2018), 1452–1462.
4. O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Res.*, **32** (2004), D267–270.
5. A. T. McCray, An upper-level ontology for the biomedical domain, *Comp. Funct. Genomics*, **4** (2003), 80–84.
6. C. G. Chute, Clinical classification and terminology: Some history and current observations, *J. Am. Med. Inform. Assoc.*, **7** (2000), 298–303.
7. N. F. Noy, D. L. McGuinness, *Ontology development 101: A guide to creating your first ontology*, Stanford Knowledge Systems Laboratory Technical Report, 2001. Available from: http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html.
8. R. Hoehndorf, P. N. Schofield, G. V. Gkoutos, The role of ontologies in biological and biomedical research: a functional perspective, *Brief. Bioinform.*, **16** (2015), 1069–1080.
9. M. Martinez-Romero, C. Jonquet, M. J. O'Connor, J. Graybeal, A. Pazos, M. A. Musen, NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation, *J. Biomed. Semantics*, **8** (2017), 21.
10. The Gene Ontology Consortium, Expansion of the Gene Ontology knowledgebase and resources, *Nucleic Acids Res.*, **45** (2017), D331–D338.
11. W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, et al., Disease Ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data, *Nucleic Acids Res.*, **43** (2015), D1071–D1078.
12. S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, et al., The Human Phenotype Ontology in 2017, *Nucleic Acids Res.*, **45** (2017), D865–D876.
13. Y. Lin, S. Mehta, H. Küçük-McGinty, J. P. Turner, D. Vidovic, M. Forlin, et al., Drug target ontology to classify and integrate drug discovery data, *J. Biomed. Semantics*, **8** (2017), 50.
14. J. Huang, K. Eilbeck, B. Smith, J. A. Blake, D. Dou, W. Huang, et al., The Non-Coding RNA Ontology (NCRO): a comprehensive resource for the unification of non-coding RNA biology, *J. Biomed. Semantics*, **7** (2016), 24.
15. N. S. Tawfik, M. R. Spruit, PreMedOnto: A Computer Assisted Ontology for Precision Medicine, in *Natural Language Processing and Information Systems* (eds. E. Métais, F. Meziane, S. Vadera, V. Sugumaran and M. Saraee), Springer, (2019), 329–336.
16. Bioportal, *Precision Medicine Ontology*, 2020. Available from: https://bioportal.bioontology.org/ontologies/PREMEDONTO/?p=classes&conceptid=root.
17. Y. He, E. Ong, J. Schaub, F. Dowd, J. F. O'toole, A. Siapos, et al., *OPMI: The Ontology of*

Precision Medicine and Investigation and its support for clinical data and metadata representation and analysis, The 10th International Conference on Biomedical Ontology (ICBO-2019), 2019. Available from: https://drive.google.com/file/d/1TN3jH4hoh40Saa8adlR_TocREGTNPVlC/view.

18. M. Uschold, M. Gruninger, Ontologies: Principles, methods and applications, *Knowl. Eng. Rev.*, **11** (1996), 93–136.
19. K. Knight, I. Chancer, M. Haines, V. Hatzivassiloglou, E. Hovy, M. Iida, et al., *Filling knowledge gaps in a broad-coverage MT system*, The 14th International Joint Conference on Artificial Intelligence, 1995. Available from: https://www.ijcai.org/Proceedings/95-2/Papers/048.pdf.
20. A. Bernaras, I. Laresgoiti, J. Corera, *Building and reusing ontologies for electrical network applications*, The 12th European Conference on Artificial Intelligence, 1996. Available from: https://www.tib.eu/en/search/id/BLCP%3ACN015300062/Building-and-Reusing-Ontologies-for-Electrical/.
21. B. Peraketh, C. Menzel, R. J. Mayer, F. Fillion, M. T. Futrell, P. S. DeWitte, et al., *Ontology Capture Method (IDEF5)*, Knowledge Based Systems, Incorporated Technical report, 1994. Available from: https://apps.dtic.mil/dtic/tr/fulltext/u2/a288442.pdf.
22. M. F. Lopez, A. Gomez-Perez, J. P. Sierra, A. P. Sierra, Building a chemical ontology using methontology and the ontology design environment, *IEEE Intell. Syst. App.*, **14** (1999), 37–46.
23. M. Gruninger, M. S. Fox, *Methodology for the design and evaluation of ontologies, Workshop on Basic Ontological Issues in Knowledge Sharing*, International Joint Conference on Artificial Intelligence, 1995. Available from: https://www.semanticscholar.org/paper/Methodology-for-the-Design-and-Evaluation-of-Gruninger/497abc0ddace6a7772a5f5a3edb3d7b751476755.
24. M. A. Musen, T. Protege, The Protege Project: A Look Back and a Look Forward, *AI Matters*, **1** (2015), 4–12.
25. M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, H. M. Abbasi, A survey of ontology learning techniques and applications, *Database (Oxford)*, **2018** (2018), bay101.
26. M. Cristani, R. Cuel, A Survey on Ontology Creation Methodologies, *Int. J. Semantic Web Inf. Syst.*, **1**(2005), 49–69.
27. National Research Council, *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*, National Academies Press, 2011.
28. National Institutes of Health, *All of Us Research Program*, 2020. Available from: https://allofus.nih.gov/.
29. M. Murphy, G. Brown, C. Wallin, T. Tatusova, K. Pruitt, T. Murphy, et al., *Gene Help: Integrated Access to Genes of Genomes in the Reference Sequence Collection*, National Center for Biotechnology Information. 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK3841/.
30. M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, et al., ClinVar: Public archive of interpretations of clinically relevant variants, *Nucleic Acids Res.*, **44** (2016), D862–D868.
31. A. T. McCray, S. Srinivasan, A. C. Browne, Lexical methods for managing variation in biomedical terminologies, *Proc. Annu. Symp. Comput. Appl. Med. Care*, **1994** (1994), 235–239.
32. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong, *Nucleic Acids Res.*, **47** (2019), D330–D338.
33. R. Winnenburg, L. Rodriguez, F. Callaghan, A. Sorbello, A. Szarfman, *Aligning pharmacologic*

classes between MeSH and ATC, International Conference on Biomedical Ontology (ICBO), 2013. Available from: http://ceur-ws.org/Vol-1061/Paper5_vdos2013.pdf.
34. QIAGEN, Relationships, 2020. Available from: http://qiagen.force.com/KnowledgeBase/KnowledgeIPAPage?id=kA41i000000L5pCCAS.
35. A. Kramer, J. Green, J. Pollard, S. Tugendreich, Causal analysis approaches in Ingenuity Pathway Analysis, *Bioinformatics*, **30** (2014), 523–530.
36. R. C. Jackson, J. P. Balhoff, E. Douglass, N. L. Harris, C. J. Mungall, J. A. Overton, ROBOT: A Tool for Automating Ontology Workflows, *BMC Bioinformatics*, **20** (2019), 407–417.
37. The OBO foundry, *Principles: Overview*, 2020. Available from: http://www.obofoundry.org/principles/fp-000-summary.html.
38. H. Sun, P. Deng, J. Li, L. Shen, Q. Qian, Automatic Concept Update Strategy Towards Heterogeneous Terminology Integration, *Data Anal. Knowl. Discov.*, **4**(2020), 121–130.
39. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nat. Biotechnol.*, **25** (2007), 1251–1255.
40. PMapp, *Chinese Program of Precision Medicine: Construction of Precision Medicine Knowledgebase for Disease Research*, 2020. Available from: http://pmap.org.cn/.
41. A. Gangemi, C. Catenacci, M. Ciaramita, J. Lehmann, *A theoretical framework for ontology evaluation and validation, Semantic Web Applications and Perspectives (SWAP)*, 2005. Available from: https://www.academia.edu/download/58656915/9.pdf.
42. P. Cimiano, J. P. McCrae, P. Buitelaar, *Lexicon Model for Ontologies: Community Report, W3C*, **2016** (2016).
43. A. Isaac, E. Summers, SKOS Simple Knowledge Organization System Reference, *W3C*, **7** (2009).

AIMS Press