Building a Knowledge Representation Model towards Precision Medicine

Institute of Medical Information & Library

Chinese Academy of Medical Sciences (CAMS&PUMC)

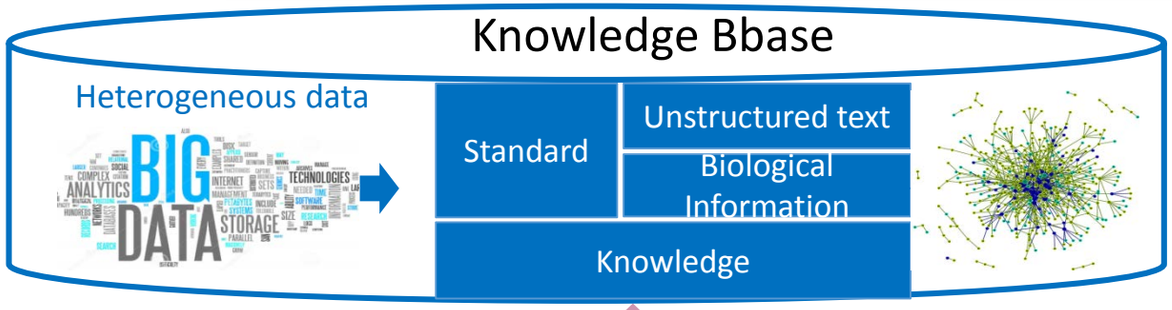
April , 2018

Outline

Background and Motivation

- The National Key Research and Development Program of China

Precision Medicine Knowledge Base



Data Integration

Knowledge Representation Model

- **Standardization**
  - Entity description
  - Relationships
  - Unique Identifiers
- **Application**
  - Text Mining
  - Data Integration
  - Knowledge Graph
  - Biocuration
  - Information Retrieval
  - Inference

What's kind of knowledge representation model?

☐ Ontology is a recognized tool to manage the heterogeneous data via defining the investigated entities and relationships between them.

- Providing identifiers for classes and relations in PM Knowledge Base , thereby enabling integration of biomedical data.

- Providing labels for classes and relations, thereby providing a Precision Medicine Vocabulary (PMV).

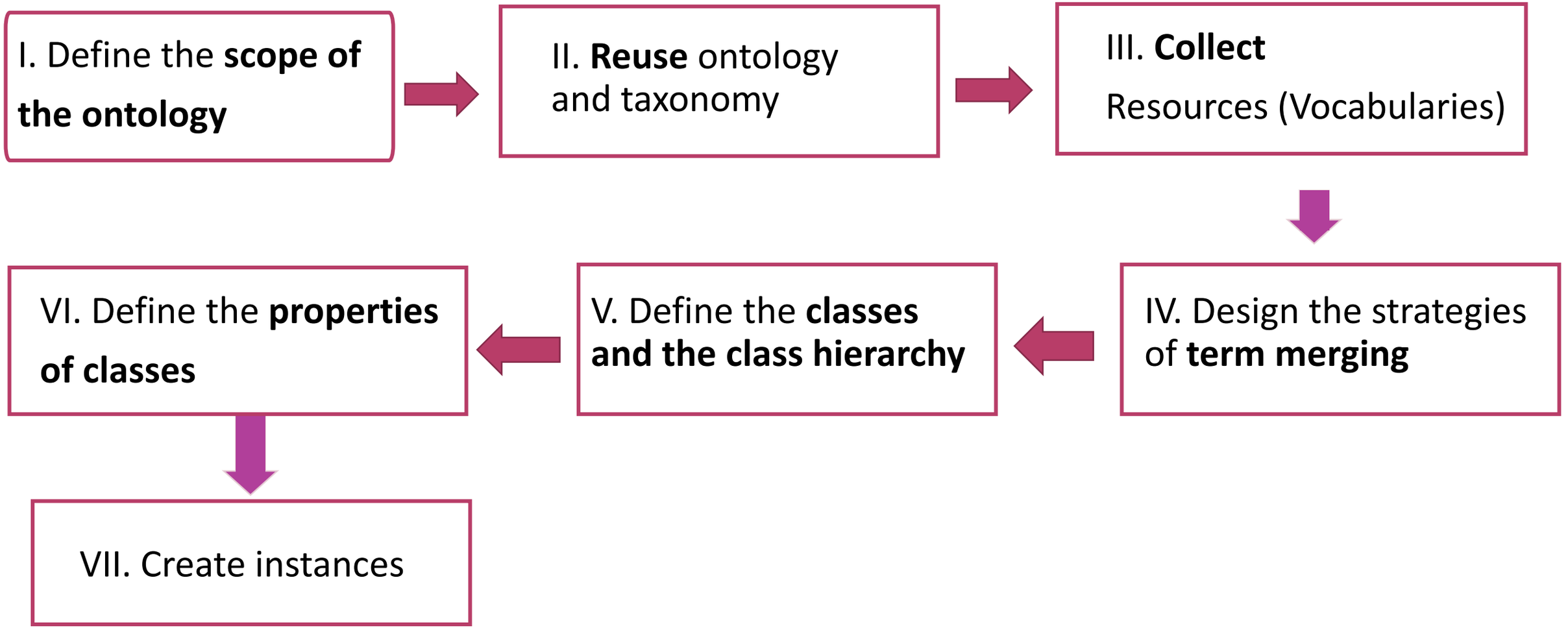
- Providing metadata associated with classes and relations that allows human users to understand their meaning and contribute to consistent use in annotation and other applications.

- Facilitating semantic integration of heterogeneous, multimodal data, and enables novel data mining in biomedical literatures.

Precision Medicine Ontology (PMO) principles

1. Covering all types of terms ,concepts and the relationship between the terms which are urgently needed by the “precision medicine project”.
2. Providing interoperability with existing international ontology .
   - Interoperable, Re-usable, Machine Readable
3. All concepts, terms and relations defined in PMO are open access
   - Findable, Open Accessible
4. Providing flexible management tools, supporting any new term merging, and new relationship created

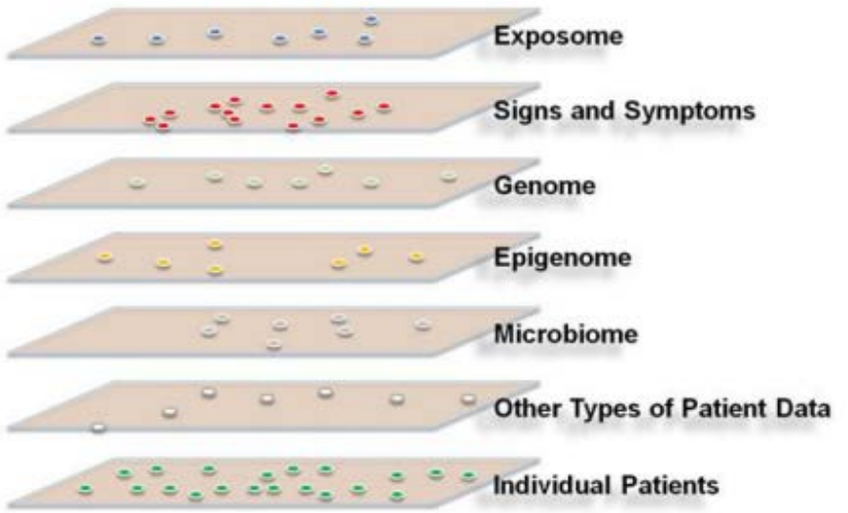
PMO construction workflow



I. Define the PMO scope

Precision Medicine

- The concept of precision medicine — prevention and treatment strategies that take individual variability into account.
- Precision medicine: take into account individual variability in genes, environment ,and lifestyle for each other.



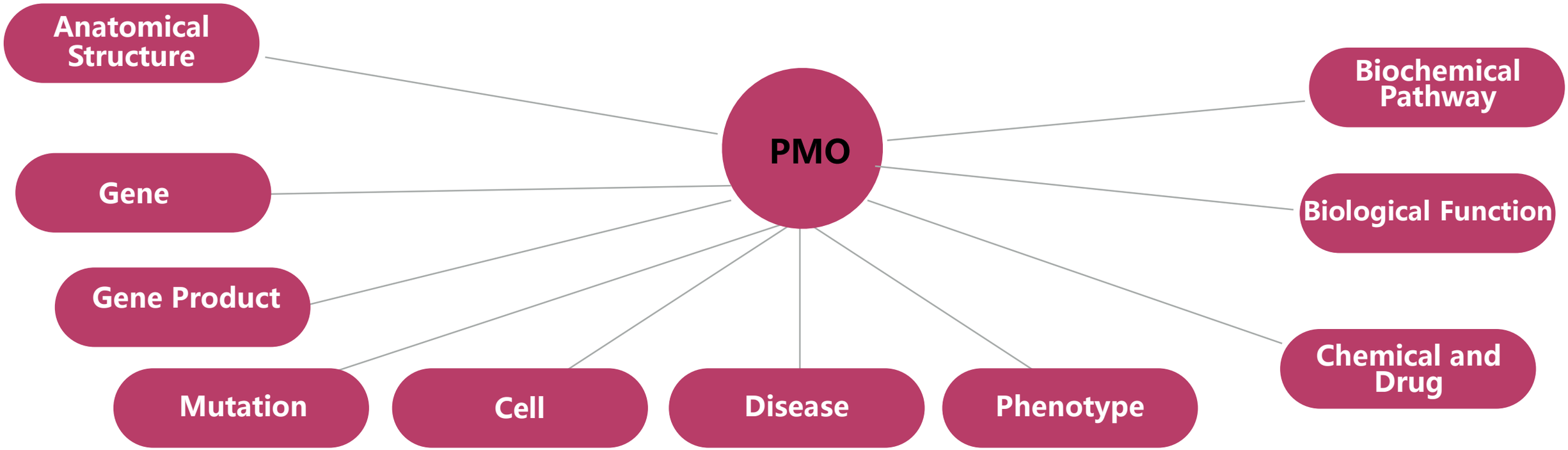
The PMO's role in the PM Knowledge Base

- Support to identify and integrate gene, disease, drug, mutation, phenotype, pathway, etc. which are the key entries in the PM knowledgebase

National Research Council. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. Washington, DC: National Academies Press, 2011

.

PMO Scope



INGENUITY IPA

NCIthesaurus 17.11

NCBI ClinVar

MeSH

NCBI Gene

50 YEARS OMIM Human Genetics Knowledge for the World

DRUGBANK Drug & Drug Target Database

Pathway Commons Over 4 000 pathways 1.3 million interactions

II. Reuse existing ontologies and taxonomy

- Diseases [C]
  - Bacterial Infections and Mycoses [C01]
  - Virus Diseases [C02]
  - Parasitic Diseases [C03]
  - Neoplasms [C04]
  - Musculoskeletal Diseases [C05]
  - Digestive System Diseases [C06]
  - Stomatognathic Diseases [C07]
  - Respiratory Tract Diseases [C08]
  - Otorhinolaryngologic Diseases [C09]
  - Nervous System Diseases [C10]
  - Eye Diseases [C11]
  - Male Urogenital Diseases [C12]
  - Female Urogenital Diseases and Pregnancy Complications [C13]
  - Cardiovascular Diseases [C14]
  - Hemic and Lymphatic Diseases [C15]
  - Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16]
  - Skin and Connective Tissue Diseases [C17]
  - Nutritional and Metabolic Diseases [C18]
  - Endocrine System Diseases [C19]
  - Immune System Diseases [C20]
  - Disorders of Environmental Origin [C21]
  - Animal Diseases [C22]
  - Pathological Conditions, Signs and Symptoms [C23]
  - Occupational Diseases [C24]
  - Chemically-Induced Disorders [C25]
  - Wounds and Injuries [C26]

MeSH

- Chemicals and Drugs [D]
  - Inorganic Chemicals [D01]
  - Organic Chemicals [D02]
  - Heterocyclic Compounds [D03]
  - Polycyclic Compounds [D04]
  - Macromolecular Substances [D05]
  - Hormones, Hormone Substitutes, and Hormone Antagonists [D06]
  - Enzymes and Coenzymes [D08]
  - Carbohydrates [D09]
  - Lipids [D10]
  - Amino Acids, Peptides, and Proteins [D12]
  - Nucleic Acids, Nucleotides, and Nucleosides [D13]
  - Complex Mixtures [D20]
  - Biological Factors [D23]
  - Biomedical and Dental Materials [D25]
  - Pharmaceutical Preparations [D26]
  - Chemical Actions and Uses [D27]

NCIthesaurus

- Biochemical Pathway
  - Cell Differentiation or Development Pathway
  - Cell Motility Pathway
  - Disease Pathway
  - DNA Repair Pathway
  - Drug Pathway
  - Genetic Information Processing Pathway
  - Metabolic Pathway
  - Regulatory Pathway
  - Signaling Pathway

- Gene Product
  - hnRNA
  - Messenger RNA
  - Precursor mRNA
  - Protein Fragment
  - Protein or Riboprotein Complex
  - Protein
  - Untranslated RNA

III. Vocabulary collection and integration

- Collecting 58 biomedical vocabularies in precision medicine domain

Gene

- Gene Ontology
- HUGO Gene Nomenclature Committee
- NCBI Gene

Mutation

- ClinVar

Disease

- ICD10
- Diagnostic and Statistical Manual of Mental Disorders
- ICPC-2 PLUS
- ……

Drug

- DrugBank
- RxNorm Vocabulary
- Anatomical Therapeutic Chemical Classification System
- Alcohol and Other Drug Thesaurus
- Gold Standard Drug Database
- FDB MedKnowledge
- Veterans Health Administration National Drug File
- ……

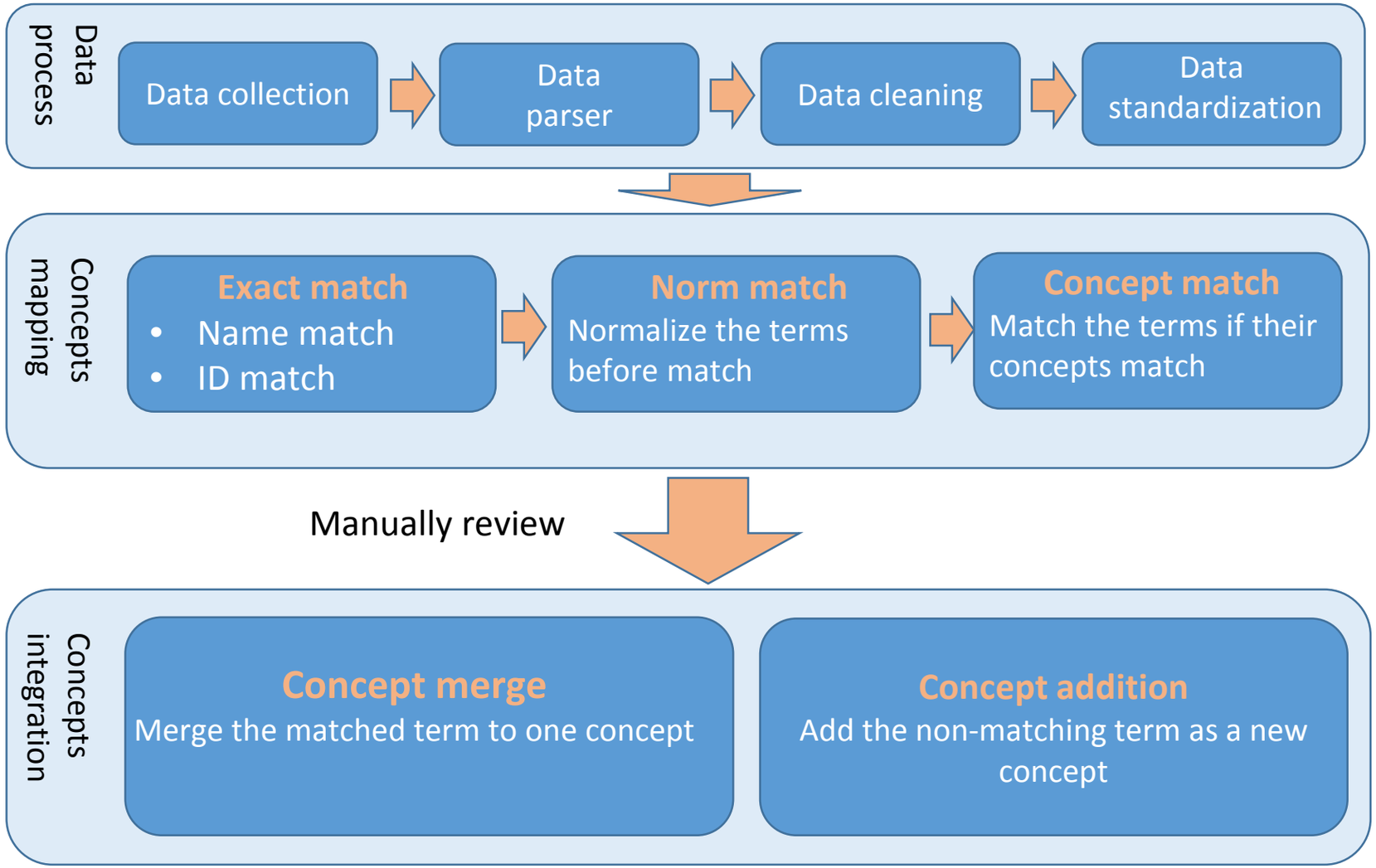
Phenotype

- Human Phenotype Ontology

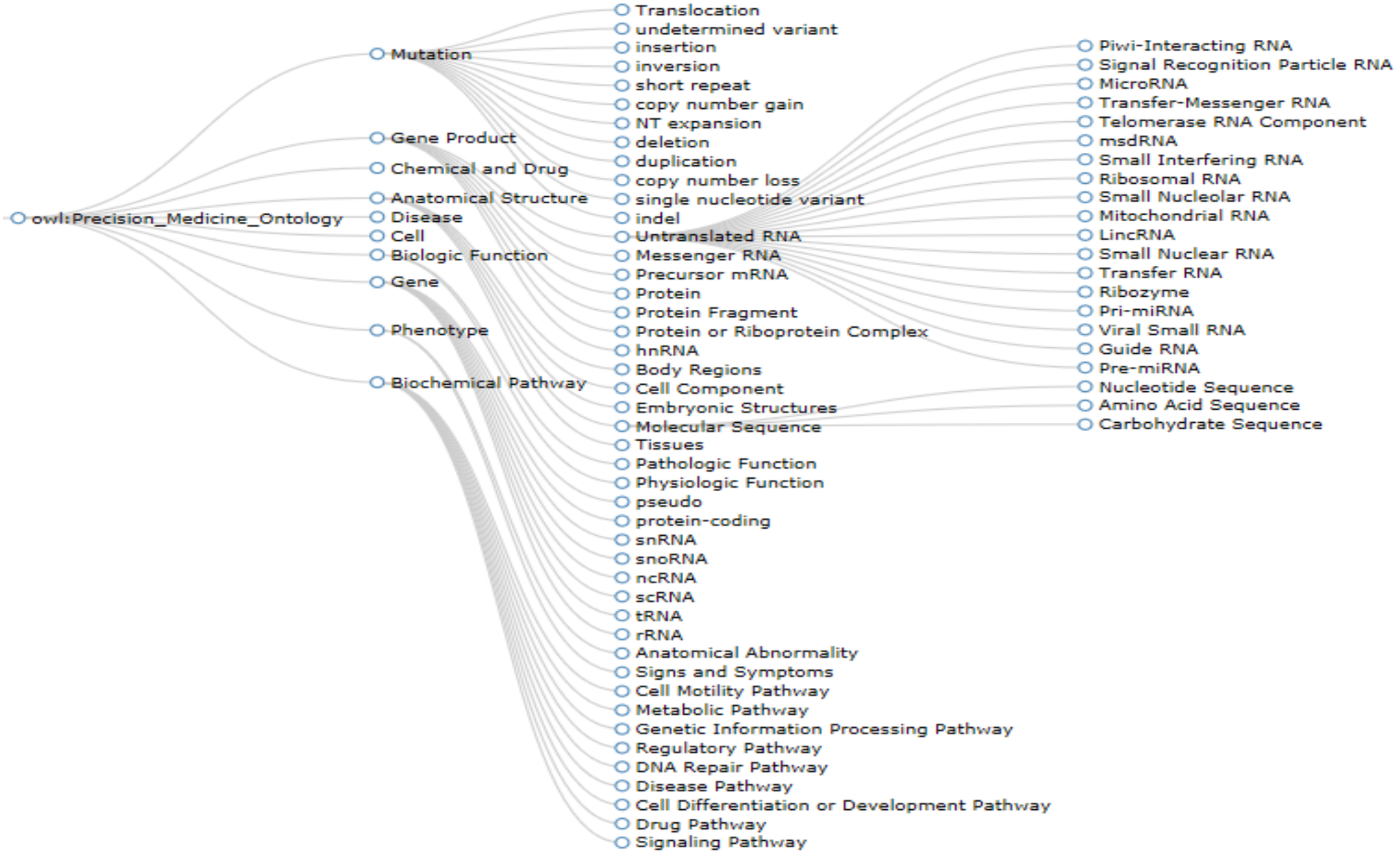
Others

- MeSH
- NCI Thesaurus
- SNOMED CT
- MedlinePlus Health Topics
- Micromedex RED BOOK
- University of Washington Digital Anatomist
- ……

IV. Term merging strategies



V. Classes and the class hierarchy

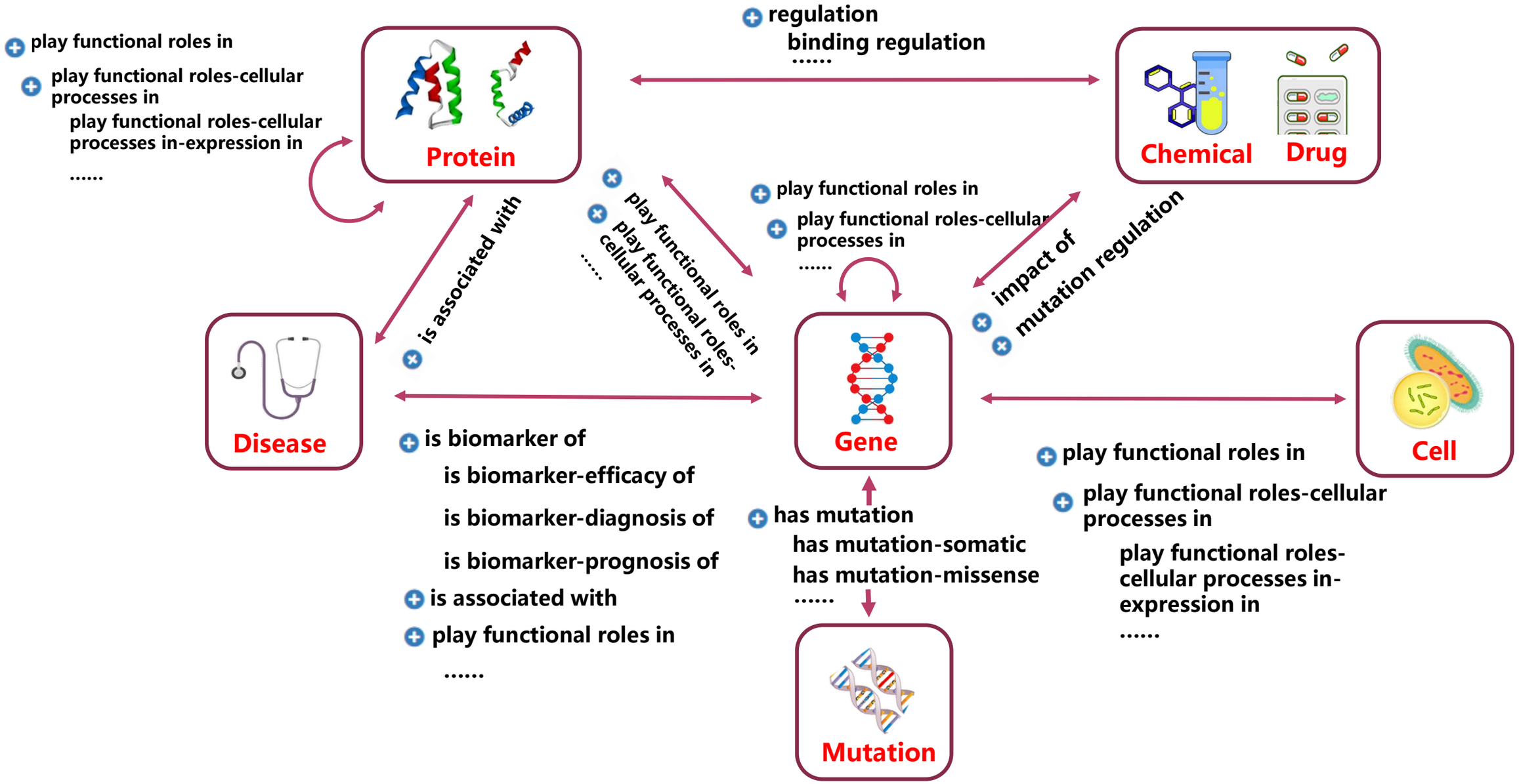


VI. Classes Properties

- Annotation properties

Annotation property	Definition
MCID	The unique identifier of the resource in PMO
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
database_cross_reference	The ID of the resource in other databases
Synonym	The synonym of the resource in other databases
is_a	The superclass of the resource
Example	The example of the relationship appearing in biomedical text
Source of Example	The source of the example, usually PubMed

Semantic Network in PMO



Semantic Relationship Definition

- Relationship Statistics: 93 relationship were organized hierarchically.

RID	Relation_Name	Domain	Range	Definition	hierarchy	Example	Source
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	Human TP53 protein has been used as a biomarker for measuring the efficacy of iloprost in the treatment of lung neoplasm.	Source: GVK Biosciences-NCT00084409
R00000003	is biomarker-diagnosis of	Gene	Disease	A gene can be used to diagnose a disease.	Second	Human TP53 gene has been used as a biomarker for diagnosis of chronic lymphocytic leukemia.	Source: GVK Biosciences-NCT00923507
R00000006	is biomarker-unspecified application of	Gene	Disease	A gene is a potential biomarker of a disease-that is, the gene (when mutated or aberrant expressed) is associated with a disease, but the precise function is unclear.	second	Upregulation of human P53 [TP53] protein in nuclei from cells malignant tumor of prostate is associated with prostate cancer in human (Biomarker status: potential biomarker).	PMID: 17145880

Summary

Conclusion

- Developing an ontology to support Precision Medicine Knowledge Base construction
- Developing a Precision Medicine Ontology (PMO):
- Defining 93 semantic relationships (such as *is biomarker of, is diagnosis biomarker of*) to connect concepts
- Making the PMO open assessable

Undergoing work

- Evaluation the PMO usability in the knowledge base construction
- Adding new classes such as Cohort and Sample
- Enriching instances in PMO from text mining and curation results (feedback)

Thanks !