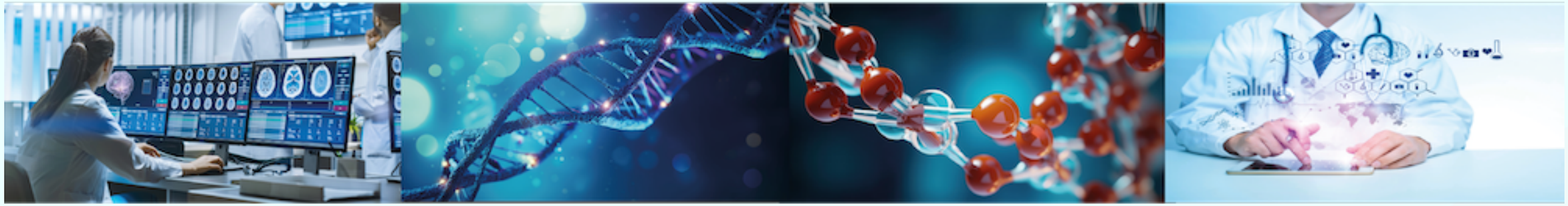


Programming for Biomedical Informatics

Lecture 15 “Structuring Biomedical Data with Ontologies”

https://github.com/tisimpson/pbi

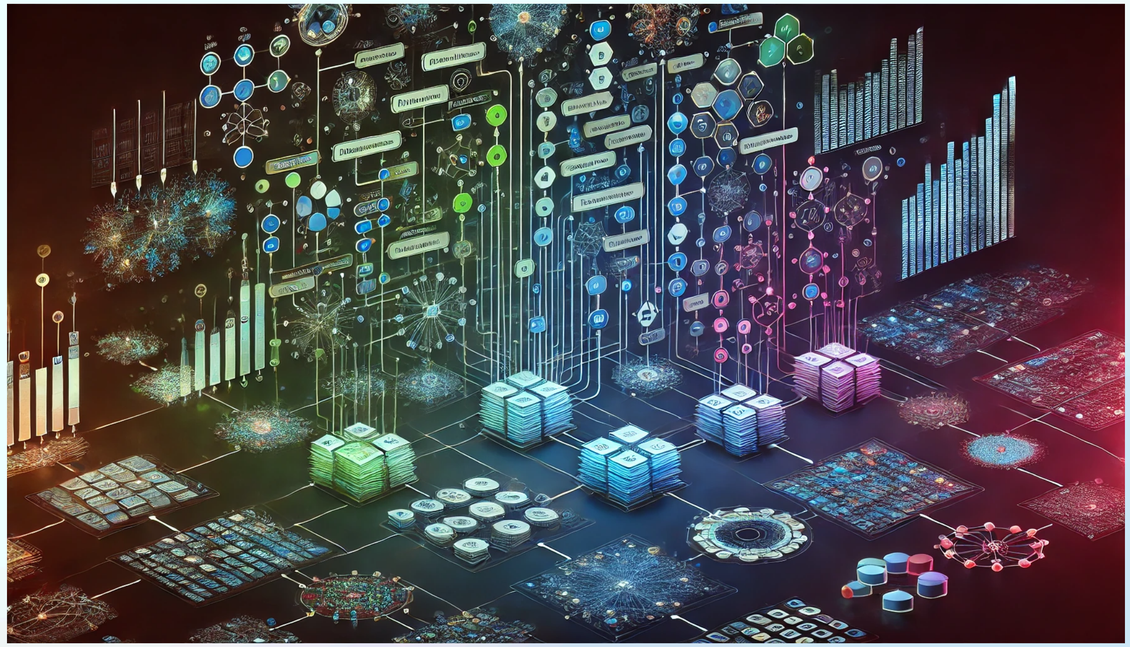
Ian Simpson
ian.simpson@ed.ac.uk

Background

THE UNIVERSITY of EDINBURGH
informatics

Biomedical & Clinical Data

- Much data in the biomedical domain is unstructured
- There are many structured data standards broadly split into terminologies and ontologies
- The quantity and heterogeneity of data is so great that manual curation into structured data objects is not feasible
- There is an unmet need to retro-fit existing data into formal data structures and to prospectively code data
- There are surprisingly few contemporary systems that do this as standard for emerging data
- Emergent properties of biological systems and their underlying mechanisms are the result of the integration of multiple biological signatures
- No single type of data can capture this
- Data integration is required that can cope with scale and complexity



Ontologies or Terminologies

Terminologies

Controlled Vocabulary
defined list of terms that are used consistently within a specific domain, ensuring standardised naming.

Synonyms and Lexical Variants
provides different names or terms that refer to the same concept, enabling matching of varied terms to the same meaning.

Hierarchical Structure
terms may be organised in a hierarchy (e.g., disease categories) with broader and narrower terms to aid in understanding relationships.

Coding Systems
often assigned alphanumeric codes (e.g., ICD-10, SNOMED CT) for easy identification and reference.

Versioning and Updates
terminologies are periodically updated to reflect advances in medicine and changes in terminology.

Cross-Referencing with Other Terminologies
terms often have mappings to other terminologies, enabling data integration across systems (e.g., mapping ICD codes to SNOMED CT).

Ontologies

Formal Representation of Knowledge
define and formalise concepts, relationships, and rules within a domain, providing a deeper semantic structure than terminologies

Conceptual Hierarchies
often have structured hierarchies that describe relationships (e.g., "is-a," "part-of") among concepts, supporting logical inferences.

Logical Axioms and Rules
define rules and logical constraints for how concepts relate, enabling reasoning engines to deduce new information.

Interoperability
are designed for integration, allowing different systems to interpret and use the same data consistently

Linking and Reuse of Concepts
often link to other ontologies (e.g., linking human disease ontology with chemical ontology), promoting data integration across disciplines.

Defined Semantic Relationships
Unlike terminologies, ontologies provide explicit definitions of relationships between terms (e.g., "causes," "treats," "affects").

Semantic Annotation
allow for tagging or annotating data with well-defined concepts, facilitating searches and analysis based on these annotations.

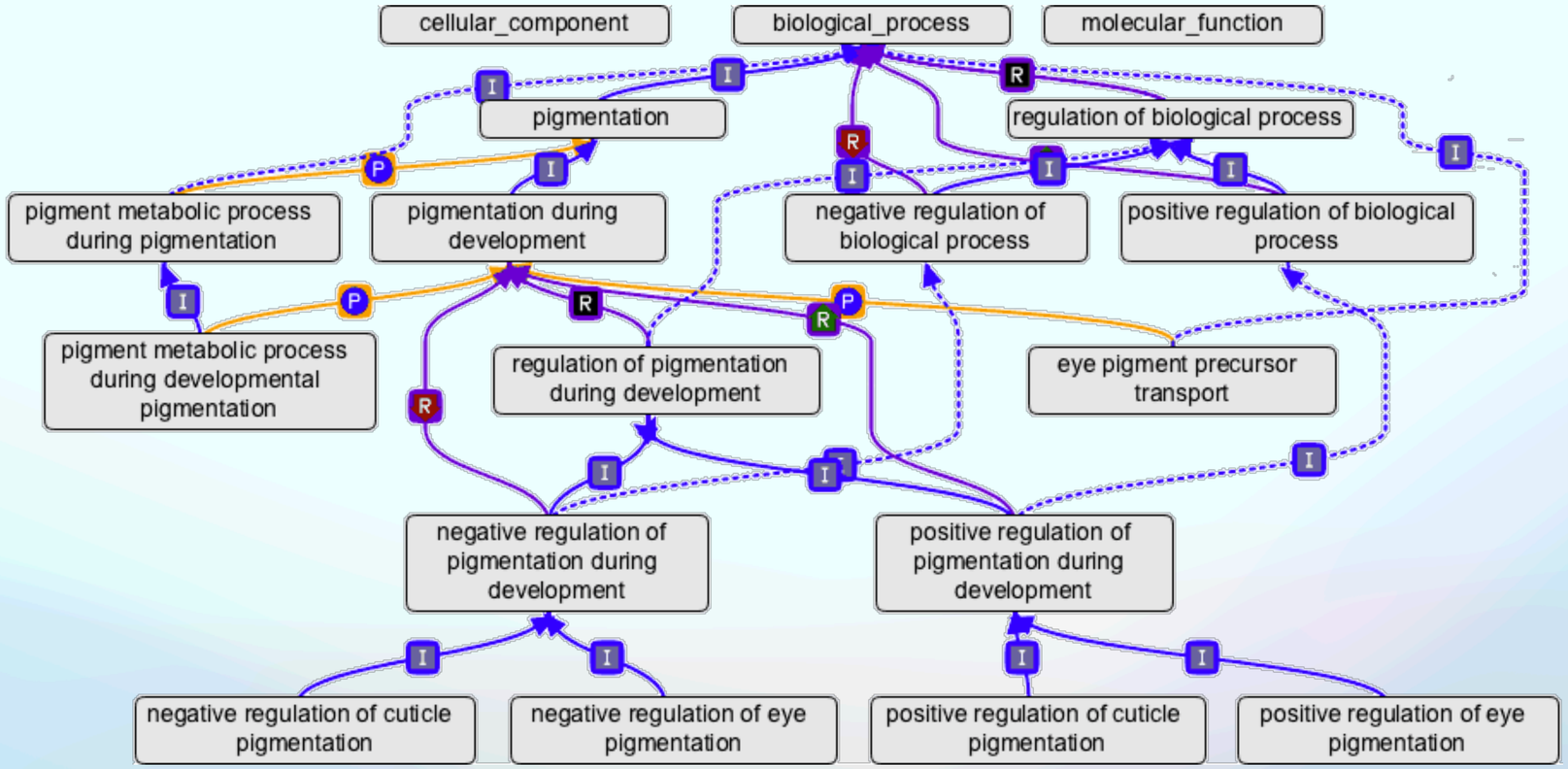
Inference Capabilities
support automated reasoning, enabling the system to infer new knowledge by applying rules to the defined relationships.

Structure of an Ontology

- nodes are “Terms”
- edges are “Relations”

low specificity roots

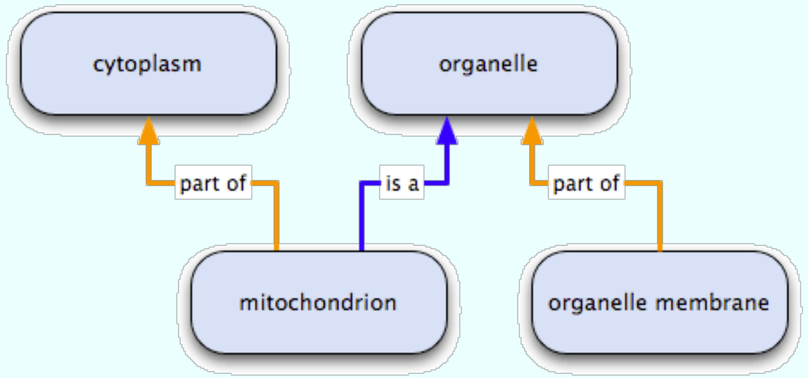
parent



child

high specificity leaves

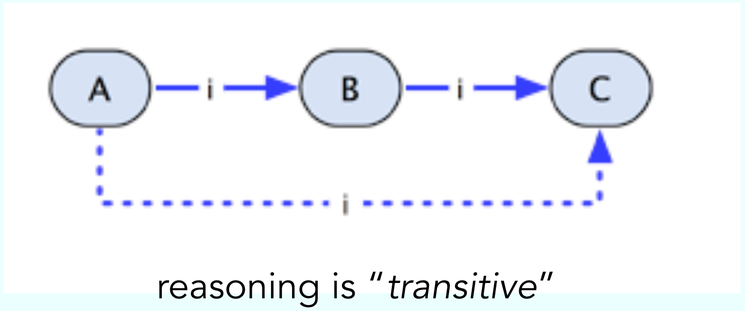
Relationships in Ontologies



parent

child

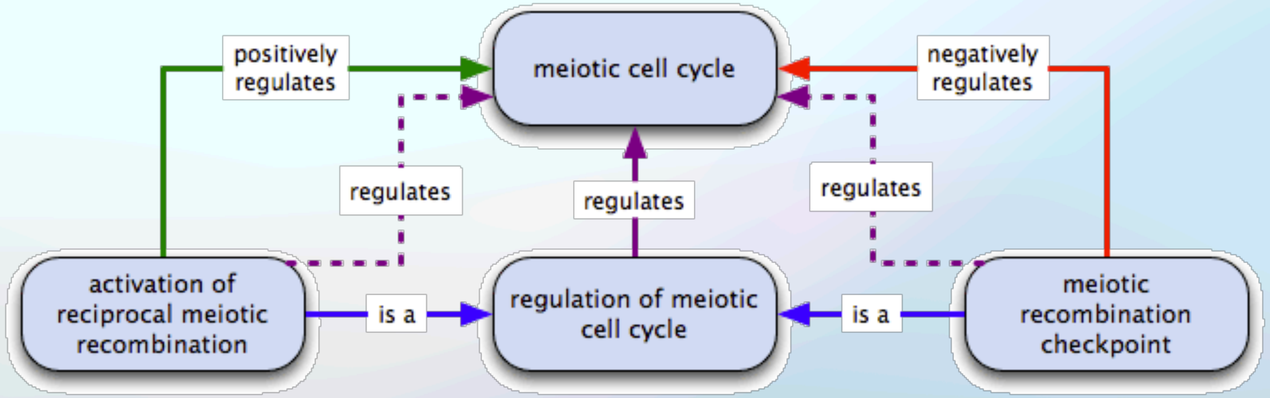
hierarchically structured



reasoning is "*transitive*"

- branching part of
- capable of
- capable of part of
- causally downstream of
- causally downstream of or within
- causally related to
- causally upstream of
- causally upstream of or within
- cell expresses
- child nucleus of
- child nucleus of in hermaphrodite
- child nucleus of in male
- coincident with
- colocalizes_with
- commensually interacts with
- composed primarily of
- concretizes
- conduit for

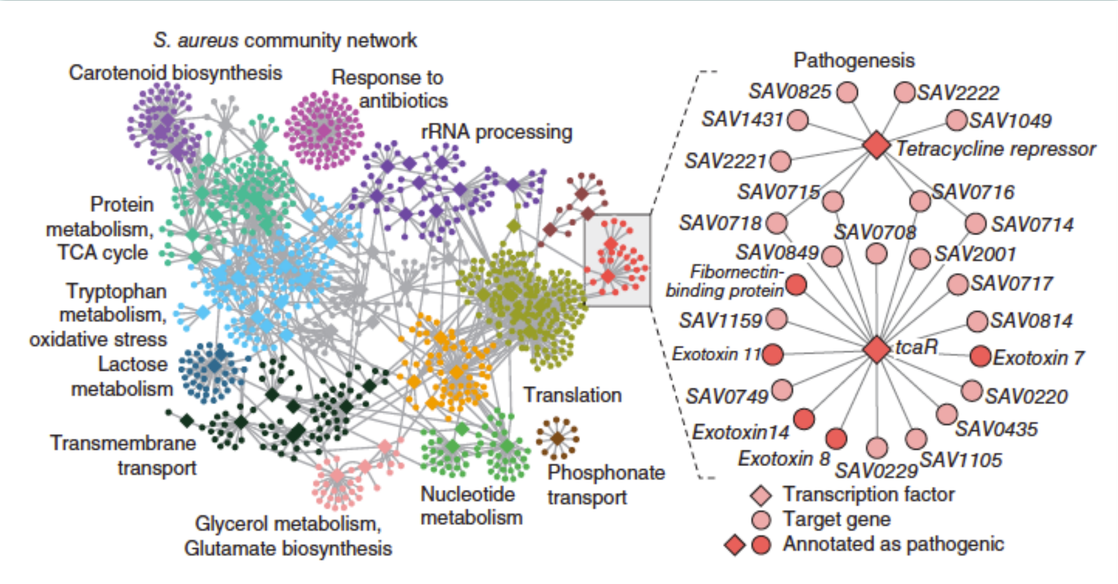
hundreds of relation types



generic AND specific relations exist

The Gene Ontology (GO)

http://geneontology.org/



Gene Ontologies

- (CC) cellular compartment
- (BP) biological process
- (MF) molecular function

Accession: GO:1901632
Name: regulation of synaptic vesicle membrane organisation
Ontology: biological_process
Synonyms
regulation of synaptic vesicle membrane organisation
regulation of synaptic vesicle membrane organisation and biogenesis
regulation of SLMV biogenesis
Definition: Any process that modulates the frequency, rate or extent of synaptic vesicle membrane organisation. Source: GOC:TermGenie, PubMed:22426000

OBO Foundry - inter-operable ontologies for bioscience

https://obofoundry.org/

Open and Collaborative
All OBO Foundry ontologies must be freely available and open for community input and collaboration.

Common Formal Language
Ontologies must be developed in a formal language, typically using OWL (Web Ontology Language), to enable computational reasoning and interoperability.

Clear and Unambiguous Definition
Each term must have a clear and objective definition, ensuring that concepts are consistently understood across applications.

Non-Overlapping Scope
Each ontology should cover a distinct domain without redundancy, minimising overlap with other OBO ontologies to prevent conflicts.

Shared and Reusable Terminology
OBO ontologies must use common, shared terms whenever possible, promoting consistency across biomedical domains.

Orthogonality and Modularity
Ontologies should be modular and orthogonal, ensuring that they can be integrated without dependency conflicts and can be independently updated.

Adherence to Best Practices
Ontologies should follow best practices in ontology development, including using standard design patterns and logical principles.

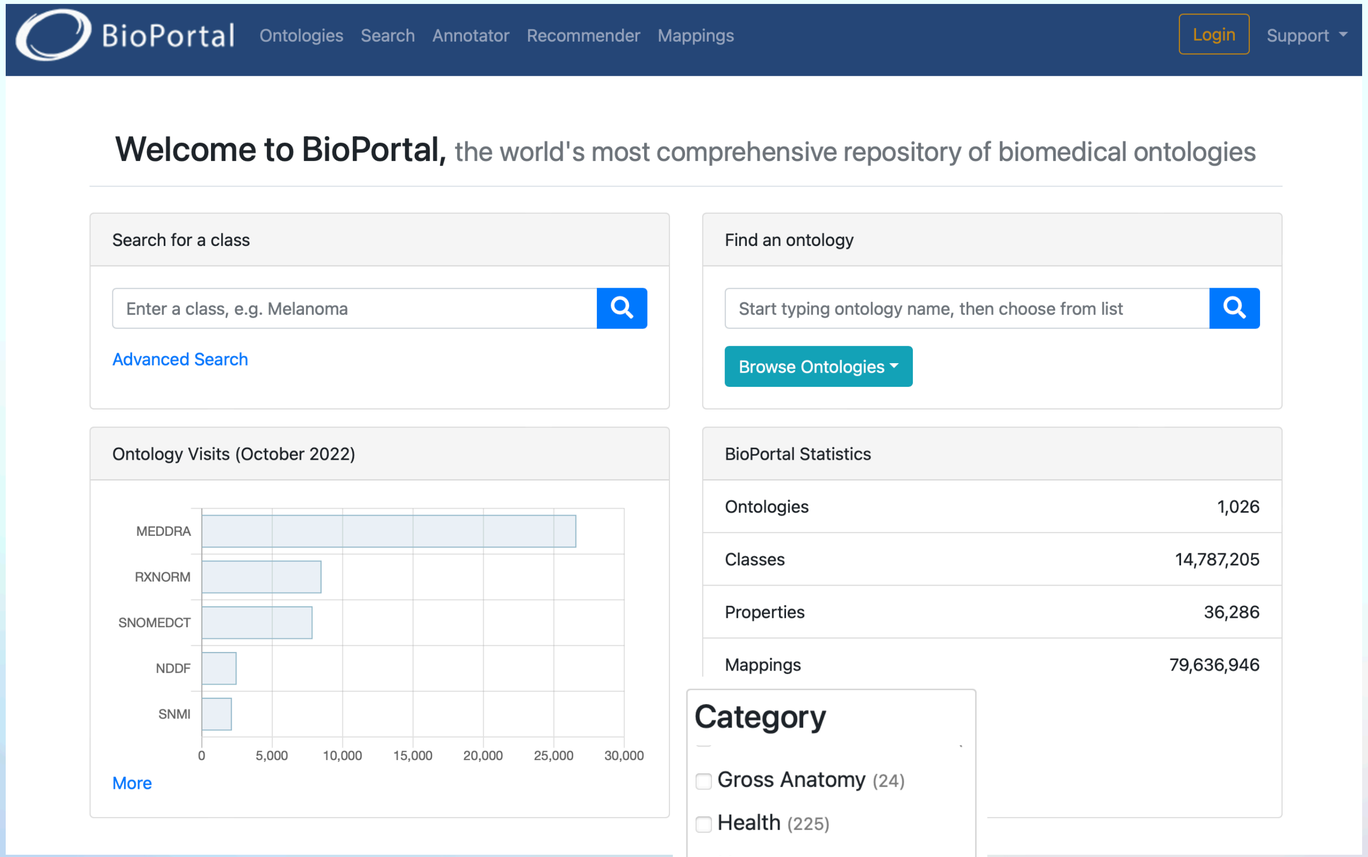
Textual Definitions and Metadata
Each concept should include a detailed, human-readable definition along with metadata for better understanding and context.

Versioning and Maintenance
Ontologies must be regularly updated, with each version documented and maintained to reflect new scientific knowledge.

Support for Community Needs
Ontologies are developed with the user community in mind, ensuring they meet the practical needs of researchers, clinicians, and developers.

BioPortal

http://bioportal.bioontology.org/



- 1157 ontologies (November 2024)
- OBO, OWL and UMLS format
- links out to original data and direct ontology download
- also available via REST
- among the most popular
  - Disease (DO)
  - Phenotype (HPO)
  - MeSH
  - OMIM
  - Taxonomical
  - Gene function (GO)
  - Anatomical

BioPortal

http://bioportal.bioontology.org/

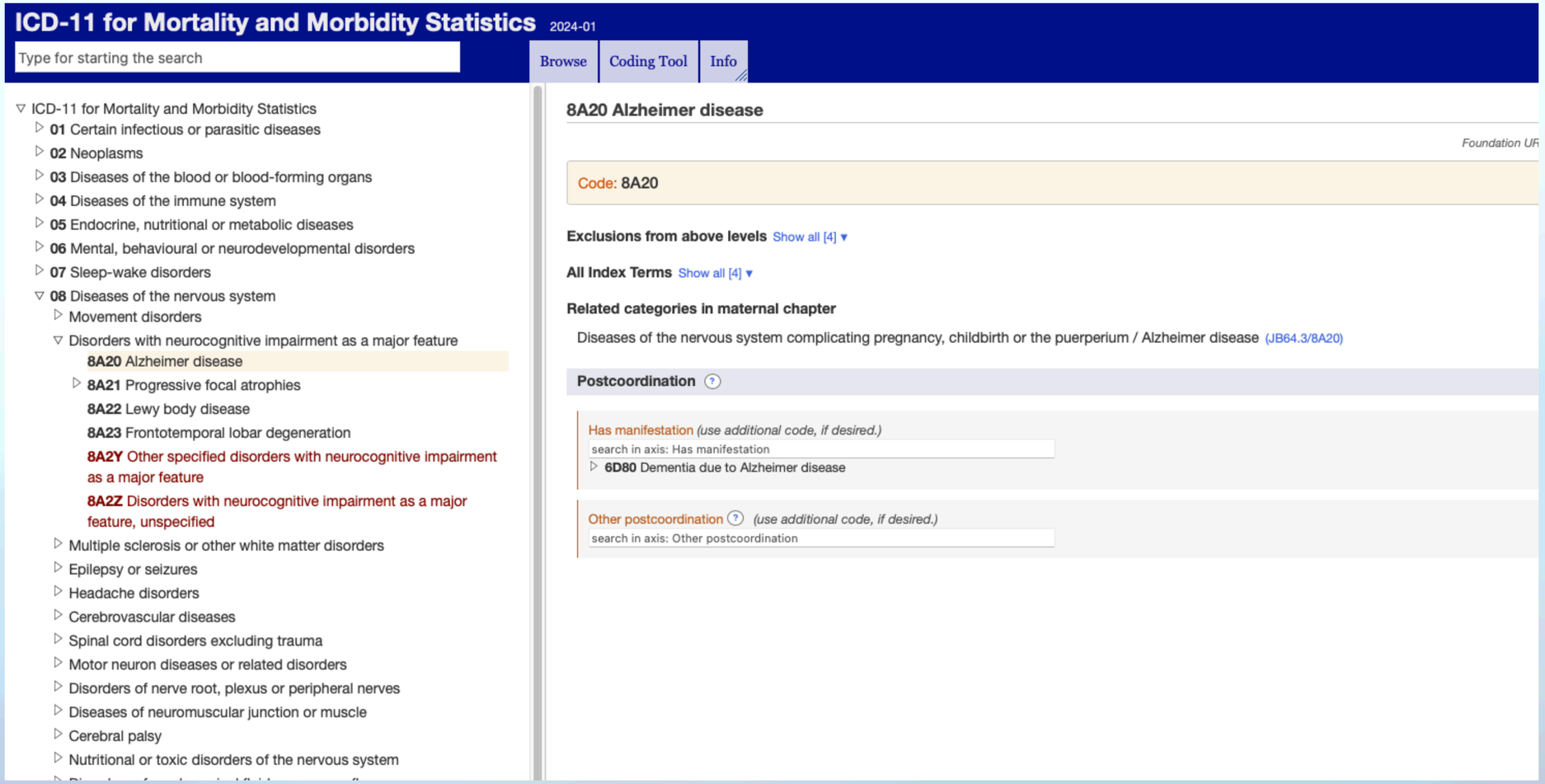
BioPortal is a comprehensive online repository that provides access to a vast array of biomedical ontologies, including many from the OBO Foundry. It aggregates and makes accessible a broader range of biomedical ontologies, including those from the OBO Foundry and many others that may not fully adhere to its principles.

- hosts ontologies from the OBO Foundry, making them easily accessible to researchers, clinicians, and developers. Users can browse, search, and download OBO Foundry ontologies via BioPortal
- includes a wider array of ontologies, including those that don’t necessarily meet all OBO Foundry criteria supporting more diverse use cases
- offers tools for exploring and mapping terms across different ontologies enhancing interoperability by allowing users to identify similar or equivalent terms across different domains
- provides metadata and version histories for ontologies, helping users track updates and changes over time
- provides visualisation tools, APIs, and services that allow users to explore the structure and relationships

Clinical Classification - ICD

https://www.who.int/standards/classifications/classification-of-diseases

The International Classification of Diseases (ICD) is a globally used diagnostic coding system developed by the World Health Organisation (WHO) to classify diseases, conditions, and health issues systematically



Clinical Classification - ICD

Alphanumeric Codes

- ICD codes are alphanumeric, with the first character typically being a letter followed by numbers
- For example, in ICD-10, codes start with a letter (A-Z) followed by two digits (e.g., A01, C34), and additional characters (like a decimal and another digit) add further specificity
  (e.g., C34.1 for specific types of lung cancer)
- ICD-10 has been superseded but is still widely used globally
- ICD-11 codes follow a similar alphanumeric pattern but have an updated format to accommodate more diseases and conditions

Hierarchical Structure

- ICD codes are organised hierarchically, with broad categories that break down into more specific subcategories
  Chapters: The highest level - each covering a broad disease category or body system e.g. Chapter II is for "Neoplasms" (C00–D48)
  Blocks: Within chapters, diseases are further organised into blocks, grouping related diseases. e.g. the neoplasms chapter includes blocks for benign, in situ, and malignant neoplasms
  Categories and Subcategories: Each block is broken down into categories and subcategories, providing increasing specificity.
  e.g. category C34 specifies malignant neoplasms of the bronchus and lung, with subcategory C34.1 for "Malignant neoplasm of upper lobe, bronchus or lung.

Coding Extensions (ICD-11)

- ICD-11 introduced “Extension Codes” for more granular details, allowing additional descriptions about the condition
  (e.g., severity, laterality, or stage of disease) without modifying the main code structure
- These “post-coordination” codes allow more detailed coding of medical scenarios by appending extensions codes to base ones

Special Characters for Additional Information

- In both ICD-10 and ICD-11, codes may include characters like a decimal point ICD-10 or slash and colon in ICD-11, e.g. E11.9 represents "Type 2 diabetes mellitus without complications”

External Causes and Other Factors

- ICD-10 and ICD-11 include chapters for coding beyond diseases, such as external causes of injuries, health status factors, social circumstances, or reasons for healthcare encounters

Clinical Terminology - SNOMED-CT

https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct

SNOMED CT (Systematised Nomenclature of Medicine – Clinical Terms) is a comprehensive, standardised healthcare terminology system used for documenting and encoding clinical data across health systems

Extensive Clinical Vocabulary
includes over 360,000 concepts, covering a wide range of clinical information such as diseases, symptoms, procedures, medications, anatomy

Hierarchical Structure
organises terms into a hierarchy, where broader concepts can be broken down into more specific terms e.g. "respiratory disorder" includes more specific terms "asthma" or "chronic obstructive pulmonary disease."

Concept-Based Approach
each concept is unique and represented by a concept ID rather than relying solely on text descriptions enabling consistent, unambiguous data representation, as concept IDs are universal and avoid issues with synonyms

Relationships and Ontological Structure
- defines relationships between concepts to enhance semantic understanding e.g. "myocardial infarction" is related to "heart disease" as an "is-a" relationship, meaning it's a type of heart disease. Other relationships like "part-of," "caused-by," or "associated with" enhance the ontology.
- supports inferencing and complex data queries, allowing clinicians and researchers to make connections between related concepts.

Support for Interoperability and Data Sharing
designed to integrate with other coding systems like ICD facilitating interoperability across health IT systems. Many EHR systems use SNOMED CT to standardise terminology, allowing healthcare data to be shared, understood, and used across different systems and regions

Cross-Mapping with Other Terminologies
provides cross-mapping with other medical coding systems like ICD-10, allowing healthcare providers to translate or map terms across systems enabling integration with billing systems, clinical research, and public health reporting

Post-Coordination
users can combine multiple concepts to capture complex or nuanced clinical scenarios. For instance, "fracture of femur" can be combined with modifiers to specify details like "left side" and "open fracture."

Clinical Terminology - SNOMED-CT

Applications of SNOMED CT

Electronic Health Records (EHRs)
used to standardise patient data, allowing precise and consistent recording of medical histories, diagnoses, treatments, and other clinical information.

Clinical Decision Support Systems (CDSS)
supports clinical decision-making by providing standardised data that can be used by algorithms to identify treatment options, flag warnings, or suggest next steps.

Public Health and Research
facilitates data aggregation and analysis, enabling researchers to identify trends, track disease outbreaks, and conduct epidemiological studies

Interoperability Across Health Systems
Common standard that enhances interoperability, enabling healthcare providers in different locations or systems to share and interpret clinical data consistently

Entry for "Diabetes mellitus type 2"

Concept ID: 44054006

Fully Specified Name (FSN)
- "Type 2 diabetes mellitus (disorder)”

Preferred Term (PT)
- "Type 2 diabetes mellitus"

Synonyms
- "Adult-onset diabetes mellitus"
- "Non-insulin dependent diabetes mellitus"
- "Type II diabetes mellitus"

Description
- "Type 2 diabetes mellitus is a chronic condition characterised by high blood glucose levels due to insulin resistance and relative insulin deficiency."

Relationships and Hierarchical Structure
- Parent Concept: Diabetes mellitus
  - Shows that "Type 2 diabetes mellitus" is a subtype of "Diabetes mellitus."
- Relationships
  - "Is a": Diabetes mellitus (this specifies that Type 2 diabetes is a type of diabetes)
  - "Associated with finding site": Pancreas structure
  - "Due to": Relative insulin deficiency
  - "Clinical course": Chronic
  - "Episodicity": Episodic
  - "Severity": Mild, Moderate, Severe (severity can be further specified)

Cross-Map to Other Coding Systems
- ICD-10 Code: E11

...

Healthcare Interoperability Standards

Fast Healthcare Interoperability Resources (FHIR)

Resource-Based, Modular Structure
organises healthcare data into reusable units called **Resources** (e.g., Patient, Observation, Medication), which can be combined to represent complex clinical scenarios.

Web-Friendly and Interoperable
uses modern web standards like RESTful APIs, JSON, and XML, allowing seamless data exchange and integration across different healthcare systems, EHRs, and patient apps.

Secure and Extensible
supports standard web security protocols for data protection and is extensible, allowing customisation to meet specific healthcare needs while maintaining compatibility with the standard.

https://fhir.org/

OMOP Common Data Model

Standardised Data Structure
provides a unified schema for healthcare data, organising information like demographics, conditions, medications, and procedures into a consistent format.

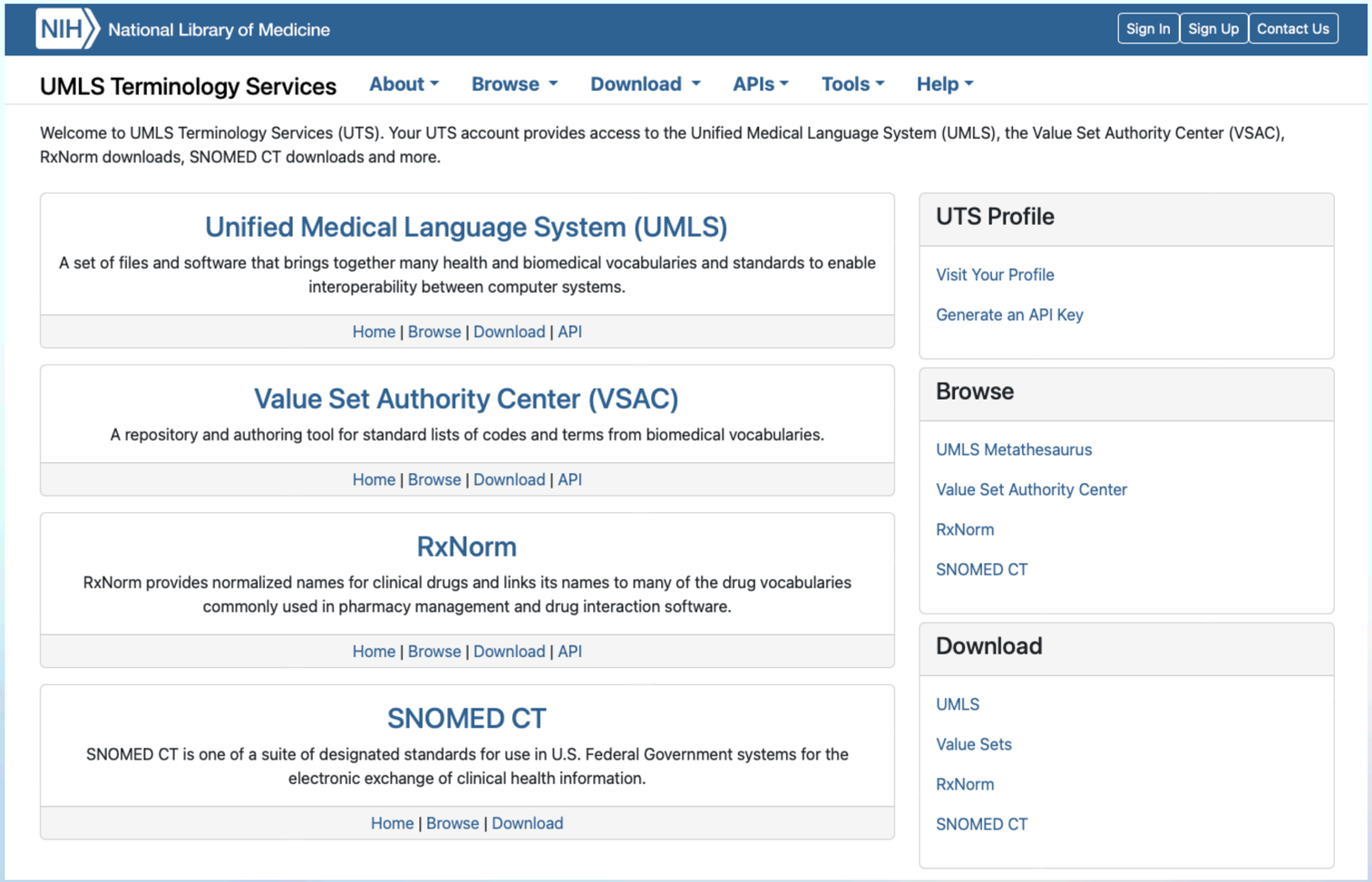
Harmonised Terminology
maps diverse coding systems (e.g., SNOMED CT, ICD) to standard vocabularies, enabling consistent terminology across datasets.

Interoperable
supports large-scale, multi-site studies and collaborative research by enabling data sharing and analysis across different healthcare systems and institutions.

https://www.ohdsi.org/data-standardization/

Unified Medical Language System UMLS

https://uts.nlm.nih.gov/



Metathesaurus: Terms and codes from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT. Hierarchies, definitions, and other relationships and attributes.

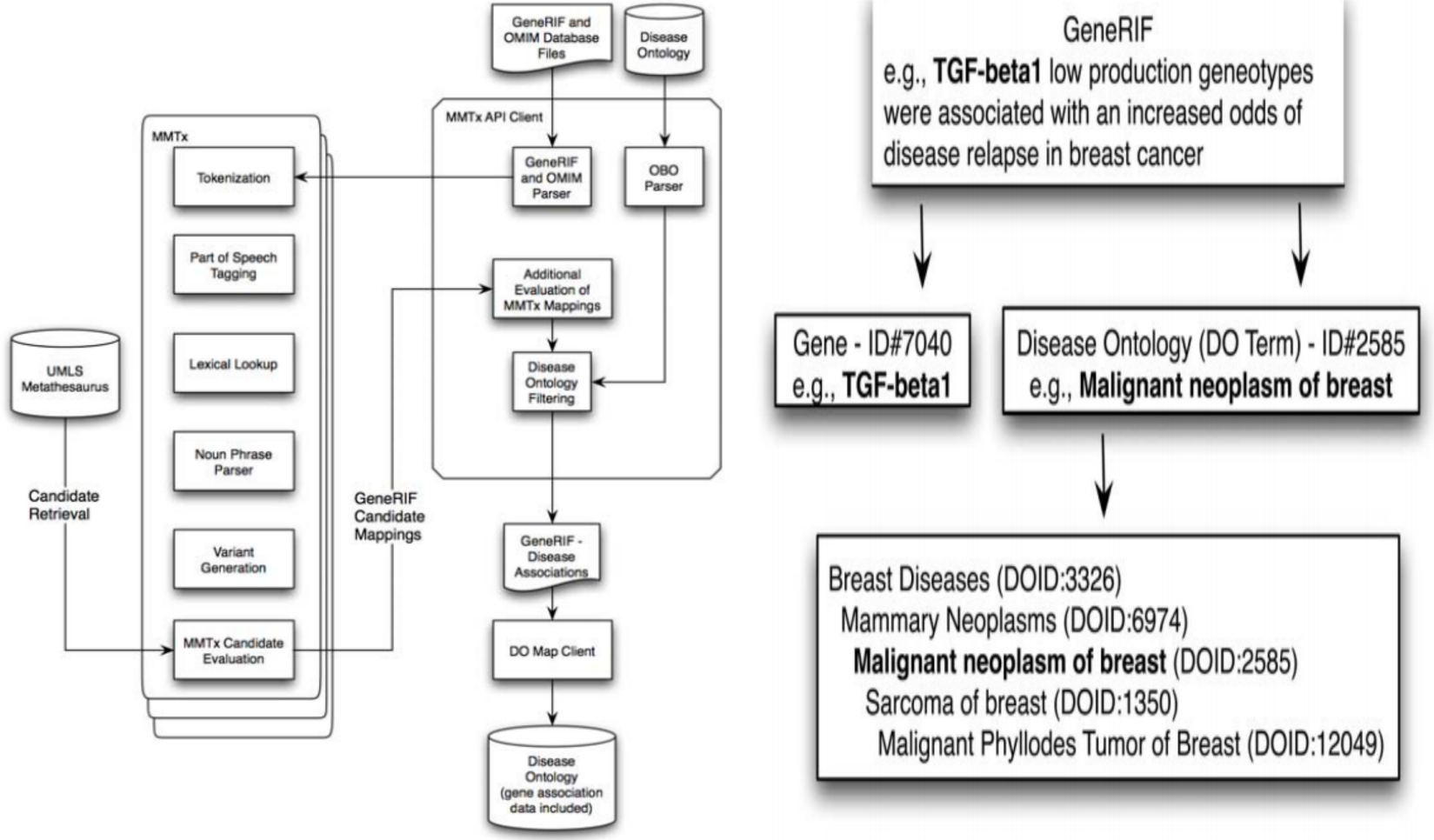
Semantic Network: Broad categories (semantic types) and their relationships (semantic relations).

SPECIALIST Lexicon and Lexical Tools: A large syntactic lexicon of biomedical and general English and tools for normalising strings, generating lexical variants, and creating indexes.

Current Release Statistics

Concepts: 3,426,877
concept names (AUIs): 16,709,195
concept names (SUIs): 13,775,220
normalised concept names (LUIs): 12,578,717
Number of sources: 170
Number of languages contributing concept names: 28

The Human Disease Ontology (DO)



18790 classes

Organised by disease category

- disease of anatomical entity
- disease of behaviour
- biological process
- environmental origin
- infectious agent and syndromes

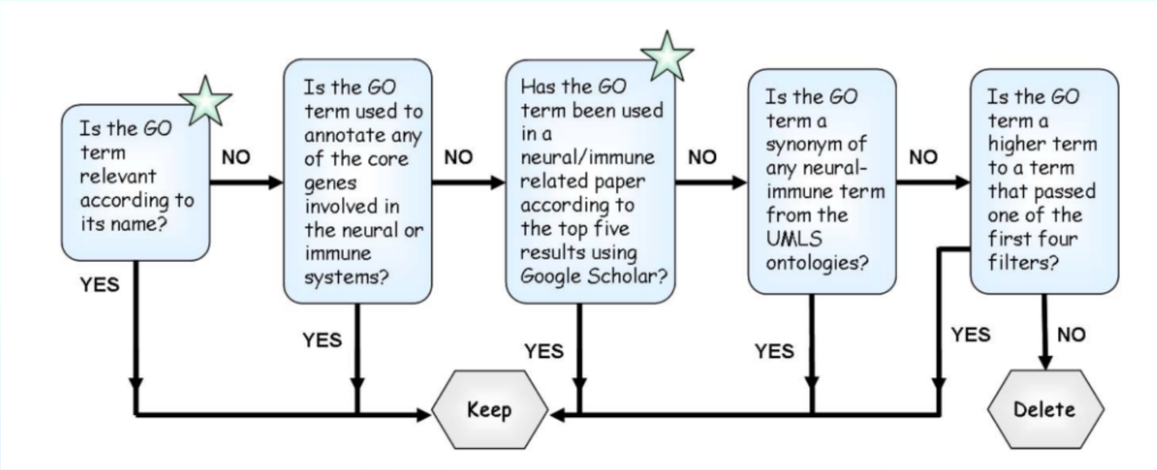
Construction

- Build evidence-based mappings phenotype and disease
- Define and validate mappings to other disease relevant
- vocabularies including
  - UMLS
  - MeSH
  - ICD
  - NCI thesaurus
  - OMIM
  - SNOMED

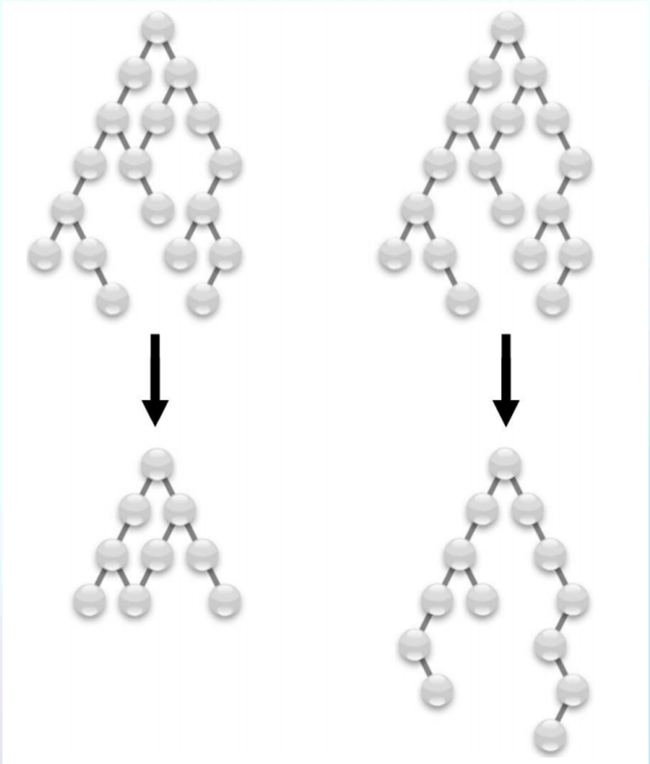
Clipped Ontologies and Slims

NiGO - the neural/immune Gene Ontology

- subset of GO directed for neurological and immunological systems
- improves statistical scores given to relevant terms
- retrieves functionally relevant terms that did not pass statistical cutoffs with full GO or the slim subset.



slimming clipping



Geifman N, Monsonego A, Rubin E. The Neural/Immune Gene Ontology: clipping the Gene Ontology for neurological and immunological systems. BMC Bioinformatics. 2010 Sep 12;11:458. doi: 10.1186/1471-2105-11-458

Tools for Working with Ontologies

Software Applications/Frameworks

Protégé
A free, open-source ontology editor developed by Stanford University.
Allows creation, editing, and visualisation of ontologies, supports OWL and RDF formats, and offers plugins for advanced functions.

https://protege.stanford.edu/

Apache Jena
A Java framework for building semantic web and linked data applications.
Supports RDF, OWL, and SPARQL queries; allows storage, manipulation, and inference of RDF data.

https://jena.apache.org/

OWL API
A Java API for creating, manipulating, and reasoning with OWL ontologies.
Supports programmatic access to ontologies, integration with reasoning engines, and compatibility with Protégé.

https://github.com/owlcs/owlapi

Python Libraries

RDFlib
- works with RDF (Resource Description Framework) data, essential for ontology manipulation
- allows parsing, serialisation, and querying of RDF data using SPARQL with RDF, RDFS, and OWL ontologies

Pronto
- allows users to load, parse, and manipulate ontologies in various formats, including OBO, OWL, and JSON.
- supports ontology operations such as accessing classes, terms, synonyms, definitions, and relationships
- designed to be simple and Pythonic with an intuitive API
- integrates with reasoning and external Libraries (like OWLready2)

OWLready2
- compatible with Protégé
- allows creating, modifying, and reasoning over OWL ontologies
- provides easy access to ontology classes, properties, and individuals

OntoSpy
- A lightweight library for inspecting and visualising ontologies.
- provides interactive exploration and visualisation of RDF, OWL, and SKOS ontologies

Literature Driven Phenotypic Gene Models

Database, 2022, 1–10
DOI: https://doi.org/10.1093/database/baac038
Original article

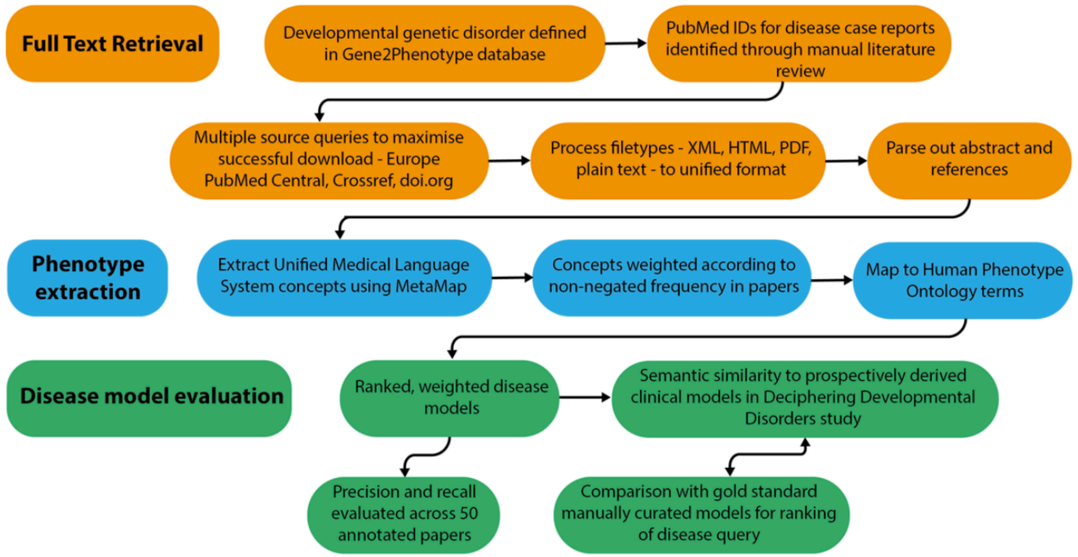
Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders

T.M. Yates[1,2], A. Lain[3], J. Campbell[1,4], D.R. FitzPatrick[1,2,4] and T.I. Simpson[3,4,*]

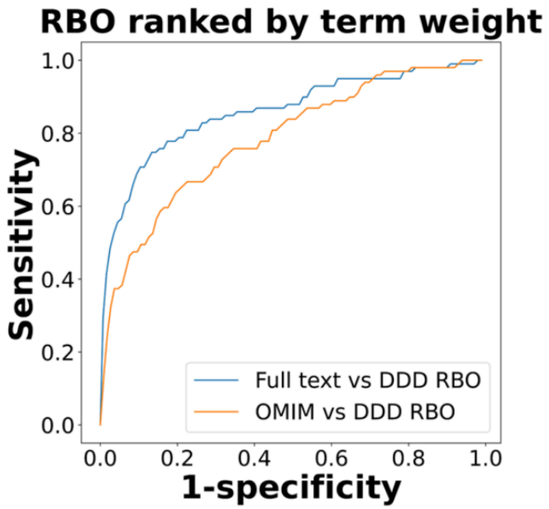
[1]MRC Human Genetics Unit, Western General Hospital, Institute of Genetics and Cancer, The University of Edinburgh, Crewe Road South, Edinburgh EH4 2XU, UK
[2]Transforming Genetic Medicine Initiative, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[3]Institute for Adaptive and Neural Computation, Informatics Forum, The University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
[4]Simons Initiative for the Developing Brain, The University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XF, UK

[*]Corresponding author: Tel: +44 (0)131 6515637; Email: ian.simpson@ed.ac.uk

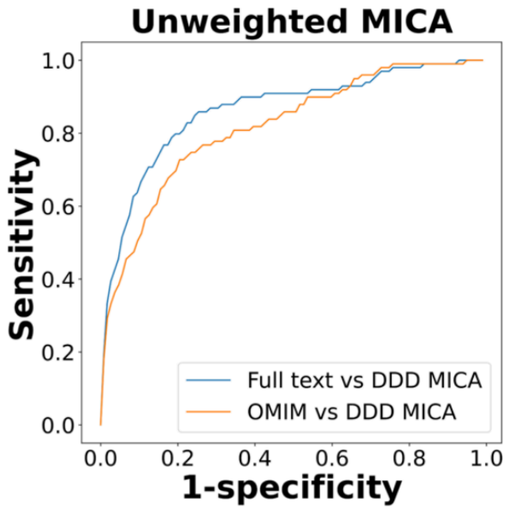
Citation details: Yates, T., Lain, A., Campbell, J. *et al.* Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders. *Database* (2022) Vol. 2022: article ID baac038; DOI: https://doi.org/10.1093/database/baac038



A



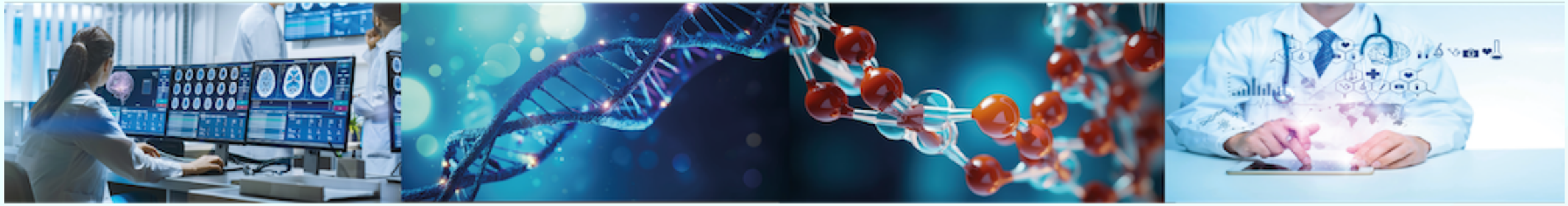
B



C

Comparator	Similarity metric	AUC
Full text vs DDD	RBO	0.850
OMIM vs DDD	RBO	0.774
Full text vs DDD	MICA	0.853
OMIM vs DDD	MICA	0.808

Yates TM, Lain A, Campbell J, FitzPatrick DR, Simpson TI. Creation and evaluation of full-text literature-derived, feature-weighted disease models of genetically determined developmental disorders. Database (Oxford). 2022 Jun 7;2022:baac038. doi: 10.1093/database/baac038



Programming for Biomedical Informatics

Next Lecture this Thursday - “Working with Ontologies & Terminologies”

Please Bring your Laptop!

Ask Questions on the EdStem Discussion Board

https://github.com/tisimpson/pbi

Coding

THE UNIVERSITY of EDINBURGH informatics