Linking Bioinformatics Research data and Publications through Metadata and Knowledge Organization Systems

Jian Qin
Syracuse University

Marcia L. Zeng
Kent State University
KOS vocabularies for representing data and publication

Facts about organisms

Knowledge derived from studying the facts about organisms

Taxonomy

Data

Subject headings
Thesauri
Classification

Publication
Example: Taxonomic representation of a DNA sequence dataset in GenBank that documents the organism in the form of taxon lineage

Example: Subject representation of the publication related to the DNA sequence dataset (PubMed ID: 7871890), which strives to provide as many and exhaustive access points as possible
Why should we care about linking data to publications?

• Evidence on which the publication is based, i.e., validity and verifiability
• Reproducibility of research
• Reuse and sharing of data more easily
Content representations for data and publications are different in terms of

1. Scope and coverage
2. Focuses or application practices
3. Ability and mechanisms for integrating biomedical research data
## 1. Scope and coverage

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Breast Neoplasms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Number</td>
<td>C04.588.180</td>
</tr>
<tr>
<td>Tree Number</td>
<td>C17.800.090.500</td>
</tr>
<tr>
<td>Annotation</td>
<td>human only; BREAST NEOPLASMS, ANIMAL EXPERIMENTAL: Manual 24.5+ neoplasm (IM)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept 1 (Preferred)</th>
<th>Breast Neoplasms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope Note</td>
<td>Tumors or cancer</td>
</tr>
<tr>
<td>Term</td>
<td>Breast Neoplasms</td>
</tr>
<tr>
<td>Term</td>
<td>Breast Tumors</td>
</tr>
<tr>
<td>Term</td>
<td>Neoplasms, Breast</td>
</tr>
<tr>
<td>Term</td>
<td>Tumors, Breast</td>
</tr>
</tbody>
</table>

| Allowable Qualifiers | BL BS CF CH CI CL CN CO DH MO NU PA PC PP PS PX RA RH |

<table>
<thead>
<tr>
<th>C50 Malig. Incl.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Excl.</td>
<td></td>
</tr>
<tr>
<td>C50.0 Nipp</td>
<td></td>
</tr>
<tr>
<td>C50.1 Cent</td>
<td></td>
</tr>
<tr>
<td>C50.2 Upper</td>
<td></td>
</tr>
<tr>
<td>C50.3 Lower</td>
<td></td>
</tr>
<tr>
<td>C50.4 Upper</td>
<td></td>
</tr>
<tr>
<td>C50.5 Lower-outer quadrant of breast</td>
<td></td>
</tr>
<tr>
<td>C50.6 Axillary tail of breast</td>
<td></td>
</tr>
<tr>
<td>C50.8 Overlapping lesion of breast</td>
<td>[See note 5 at the beginning of this chapter]</td>
</tr>
<tr>
<td>C50.9 Breast, unspecified</td>
<td></td>
</tr>
</tbody>
</table>

Source: [http://apps.who.int/classifications/icd10/browse/2016/en](http://apps.who.int/classifications/icd10/browse/2016/en)

Image credit: [http://www.physio-pedia.com/Physiotherapy_and_cancer_treatment](http://www.physio-pedia.com/Physiotherapy_and_cancer_treatment)
Constraints of conventional KOS vocabularies on coverage and scope

Coarse granularity on representing concepts and relationships

Covert relationships between concepts

Documentation of information about a concept
2. Focuses or application practices

Organizing knowledge of organisms

- Applying scientific taxonomy and nomenclature to
  - identify, 
  - name, and 
  - classify them 
- in bioinformatics data, and 
- in the metadata that describes such data.

Organizing information and knowledge contained in research publications

- Applying thesauri and classifications to
  - index,
  - retrieve,
  - organize, and
  - connect
- the scholarly output from studying the organisms, and
- the scholarly output in regulation and guideline documents.

Examples
- *NCBI* Taxonomy
- *NCBI Organismal Classification*

*NCBI = National Center for Biological Information

*NCI = National Cancer Institute

### 3. Ability and mechanisms for integrating biomedical research data

*NCBI Organismal Classification*

*NCBI = National Center for Biological Information*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxonomy ID</td>
<td>11103</td>
</tr>
<tr>
<td>Inherited blast name</td>
<td>viruses</td>
</tr>
<tr>
<td>Rank</td>
<td>species</td>
</tr>
<tr>
<td>Genetic code</td>
<td>Translation table 1 (Standard)</td>
</tr>
<tr>
<td>Host</td>
<td>vertebrates human</td>
</tr>
<tr>
<td>Lineage( full )</td>
<td>Hepatitis C virus</td>
</tr>
<tr>
<td><em>Lineage</em></td>
<td>Hepatitis C</td>
</tr>
</tbody>
</table>

#### Genome Information

| Source: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&name=Hepatitis+C+virus&lvl=0&srchmode=1 |

### The Genetic Codes

#### 1. The Standard Code (transl_table=1)

By default all transl_table in GenBank flatfiles are equal to id 1, and

- TTT F Phe
- TCT S Ser
- TAT Y Tyr
- TGT C Cys
- TTC R Arg
- TAC Y Tyr
- TCC C Cys
- TGG W Trp
- TCT C Cys
- TGC R Arg
- TCA S Ser
- CGA R Arg
- CGG R Arg
- GGT V Val
- GGC R Arg
- GCA S Ser
- GGG G Gly
- AAA K Lys
- AAG K Lys
- GAT S Ser
- GAA G Gly
- GTT V Val
- GGT V Val
- GAG G Glu
- AGT S Ser
- GGT V Val
- TCT C Cys
- TGG W Trp
- TTC R Arg
- TAC Y Tyr
- TCC C Cys
- TGG W Trp
- TCT C Cys
- TGG W Trp
- TTC R Arg
- TAC Y Tyr
- TCC C Cys
- TGG W Trp
- TCT C Cys
- TGG W Trp

3. Ability and mechanisms for integrating biomedical research data

**NCI Thesaurus (NCIt)**

*NCI = National Cancer Institute*

**Hepatitis C Virus (Code C14312)**

Table of Contents
- Terms & Properties
- Synonym Details
- Relationships
- Mappings

Terms & Properties
- Preferred Name: Hepatitis C Virus
- Definition: A small, enveloped, positive sense single strand RNA virus in the family flaviviridae.
- CDISC Definition: Any viral organism that can be assigned to the species Hepatitis C virus.
- Label: Hepatitis C Virus
- NCI Thesaurus Code: C14312 (Search for linked caDSR metadata) (see value)
- NCI Metathesaurus Link: C0220847 (see NCI Metathesaurus info)
- Synonyms & Abbreviations: (see Synonym Details)
- HCV, Hepatitis C, HEPATITIS C VIRUS, Hepatitis C Virus, Virus-Hepatitis C
- External Source Codes: SNOMED CUI C0220847

Synonym Details
- Term | Source | Type | Code
- HCV, Hepatitis C virus | NCI | AB | C14312
- HCV | CDISC | SY | C14312
- Hepatitis C | NCI | SY | C14312
- HEPATITIS C VIRUS | CDISC | PT | C14312
- Hepatitis C Virus | NCI | PT | C14312
- Virus-Hepatitis C | NCI | SY | C14312

**Relationships with other NCI Thesaurus Concepts**

Parent Concepts:
- Hepacivirus
- Hepatitis Virus

Child Concepts: (none)

Role Relationships pointing from the current concept to other concepts: (none)

Associations pointing from the current concept to other concepts:
- (True for the current concept.)
  - Concept_In_Subset: CDISC SDTM Microorganism Terminology
  - Concept_In_Subset: CDISC SDTM Species Terminology
  - Concept_In_Subset: CDISC SMTM Terminology

**Hepatitis C Virus (CUI C0220847)**

Incoming Role Relationships pointing from other:

Incoming Associations pointing from other concepts:

Mapping relationships:
- see Mappings

Mapping Details
- Hepatitis C Virus C0220847

https://ncit.nci.nih.gov/ncitbrowser/pages/concept_details.jsf?dictionary=NCI_Thesaurus&version=16.09d&code=C14312&ns=NCI_Thesaurus&type=all&key=n1875063326&b=1&n=0&vse=null
Ways of linking data to publication

• Identifiers
  - DOI
  - URI

• Semantic relationships
  - KOS crosswalk
  - Co-indexing terms
  - Knowledge networks

- Object-to-object linking
- Concept-to-concept linking
- Label-to-term linking
- Node-to-node linking
Object-to-object linking

A publication ID from PubMed is embedded in the dataset’s metadata record.
Many KOS vocabularies already exist, maybe mapped...

Concept-to-concept linking
Metadata for a DNA sequence dataset in the GenBank data repository

**LOCUS**  SCU49845  5028 bp  DNA

**DEFINITION**  Saccharomyces cerevisiae TCP1-beta gene (AXL2) and Rev7p (REV7) genes, complete

**ACCESSION**  U49845

**VERSION**  U49845.1  GI:1293613

**SOURCE**  Saccharomyces cerevisiae (baker's yeast)

**ORGANISM**  Saccharomyces cerevisiae

Eukaryota; Fungi; Ascomycota; Saccharomycetales; Saccharomycetaceae

**MeSH Terms**
- Amino Acid Sequence
- Base Sequence
- Chromosomes, Fungal
- Cloning, Molecular
- DNA Damage*
- DNA Replication
- DNA, Fungal/biosynthesis
- DNA, Fungal/secretion
- DNA-Directed DNA Polymerase*
- Fungal Proteins/chemistry
- Fungal Proteins/genetics*
- Genes, Fungal*
- Genetic Complementation Test
- Molecular Sequence Data
- Mutagenesis*
- Open Reading Frames

**Label-to-term linking**

Indexing terms in PubMed for the paper that resulted from studying the DNA sequence

**Substances**
- DNA, Fungal
- Fungal Proteins
- REV7 protein, S cerevisiae
- Saccharomyces cerevisiae Proteins
- DNA-Directed DNA Polymerase
Node-to-node linking

(Source: NAS, 2011)
The next question is: How are we going to create the links?
A case study: Knowledge nodes in precision medicine publications
The vision of a Knowledge Network of Disease and Information Commons

(Source: NAS, 2011)
Research problem

“Because new information and concepts from biomedical research cannot be optimally incorporated into the disease taxonomy of today, opportunities to define diseases more precisely and to inform health-care decisions are being missed.”

(Source: NAS, 2011)
Approach to address the research problem

New information and discovery is reported in Knowledge nodes:
- Types?
- Attributes?

Identify from publications

Attributes of data (metadata)

Object-to-object linking
- Concept-to-concept linking
- Label-to-term linking
- Node-to-node linking

Knowledge nodes:
- Types?
- Attributes?

Relationships between nodes:
- Types?
- Attributes?
Pilot study: data

- A sample of 30 articles in precision medicine
  - Four in breast cancer
  - Five in diabetes
  - Eleven in oncology
- “Purposeful sampling”
  - To gain insights and in-depth understanding rather than empirical generalizations
Pilot study: Selecting knowledge nodes

- Molecular entities such as genes, proteins, genomes, etc.
- Disease names
- Names or terms related to treatments/therapies
- Methods, techniques, and types of decisions related to diagnosis
- Data sources used by the publication
- Types of relationships between potential knowledge nodes
Pilot study: Mapping knowledge nodes

A sample map of knowledge nodes and relationships from a research paper (based on PubMed paper ID 25441102)
Pilot study: Preliminary results (1)

- Structural levels of nodes

Examples of knowledge nodes derived from the sample publications

<table>
<thead>
<tr>
<th>Category</th>
<th>Atomic level (name of things)</th>
<th>Concept level</th>
<th>Cluster level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>Her2, BRCA1, BRCA2, EGFR</td>
<td>Oncogenes</td>
<td>EGFR mutations in lung cancer</td>
</tr>
<tr>
<td>Disease</td>
<td>Non-squamous carcinoma, squamous cell carcinoma</td>
<td>Non-small cell lung cancer</td>
<td>Lung cancer</td>
</tr>
<tr>
<td>Drug</td>
<td>Pertumzumab, Lmatinib, Crizotinib</td>
<td>Tyrosine kinase inhibitor</td>
<td>Oncogene de-addiction</td>
</tr>
</tbody>
</table>
Pilot study: Preliminary results (2)

• Knowledge nodes by
  – Disciplinary field:
    genetics, pathology, pathophysiology, oncology, virology, ...
  – Disease name and biomarker pairs:
    • Chronic myeloid leukemia (CML) with mutated gene BCR-ABL
    • Breast cancer with positive estrogen receptor (ER), BRCA1/2, and Her2
    • Non-small cell lung cancer with mutations in multiple genes such as epidermal growth factor receptor (EGFR), excision repair-cross complementation group (ERCC), and ribonucleotide reductase (RRM)
Pilot study: Preliminary results (3)

- Knowledge nodes that blend clinical and basic research
  - clinically actionable mutations
  - phenotype of breast cancer
  - resistance to endocrine therapy
  - biomarkers predicting response to therapy
  - genomic drivers of cancer
  - predictive and prognostic biomarkers
  - intratumor heterogeneity
  - molecular classification of tumors
Pilot study: Preliminary results (4)

Major relationships types and patterns between knowledge nodes observed in the sample publications

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>has-biomarker</td>
<td>Disease has-biomarker Gene</td>
<td>chronic myeloid leukemia has-biomarker BCR-ABL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>non-small cell lung cancer has-biomarker EGFR</td>
</tr>
<tr>
<td>is-driver-of</td>
<td>Gene is-driver-of Disease</td>
<td>Her2 is-driver-of breast cancer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c-Kit is-driver-of chronic granulocytic leukemia</td>
</tr>
<tr>
<td>targets</td>
<td>Drug targets Gene</td>
<td>Crizotinib targets ALK</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Olaparib targets BRCA1/2</td>
</tr>
<tr>
<td>has-role-of</td>
<td>Drug has-role-of Treatment</td>
<td>Crizotinib has-role-of oncogene de-addiction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Olaparib has-role-of DNA repair</td>
</tr>
</tbody>
</table>
Implications of preliminary results

• Knowledge nodes may be marked with different labels—structure, discipline, disease, gene or biomarker, treatment, ...
• Each label represents a dimension and the nodes in one dimension form a vector
• A node may reside in multiple dimensions at the same time
• The knowledge network of disease can be considered as the sum of nodes in all vectors, which becomes a data science research problem
Concluding remarks

• Linking between data and publications requires reexamining the data and knowledge landscape and renew our understanding of KOS in the context of data-intensive science
• New types of KOS need to be dynamic, flexible, and linkable
• Models, patterns, and computational algorithms will be needed to develop the knowledge network of disease that incorporates basic science with clinical practice
References