What to think about when choosing and building on vocabularies for your work

Margaret W. Haber
Associate Director
NCI Enterprise Vocabulary Services
October 2012
Why is Terminology Important?

- Terminologies provide the underlying foundation upon which much of the electronic sharing and re-use of data, services and applications depend. NIH Institutes and collaborating organizations such as FDA and CDISC all require the development and use of this fundamental resource.

- Terminology is also the primary foundation upon which metadata and data model infrastructure are built. Understanding the semantics or essential meaning of a domain or content area requires that any models, metadata, value sets, or other artifacts such as case report forms use consistent, well-formed terminologies for their annotation, definition and population.

- In providing controlled terminology over the past 15 years, NCI Enterprise Vocabulary services (EVS) has found that many needs are shared with other ICs, agencies, SDOs, and researchers, so forging partnerships for development both spreads the load and builds better standards.
Terminology Provides Consistent Meaning
http://www.nichd.nih.gov/health/clinicalresearch/terminology/about.cfm
Terminology Usage: Diversity

Terminologies are usually created for and reflect a particular purpose. In meeting one purpose well, they often have trouble meeting others.

- The ICDs were originally designed for morbidity coding, but have severe limitations for accurately describing patients or coding research activities.
- MeSH is uniquely designed to optimize relevant literature retrieval, but is not designed for clinical coding.
- MedDRA is optimized for adverse event reporting, but not for diagnosis, etc.

We can expect that it will be necessary to have a range of vocabularies for particular purposes well into the future.
NCI Term Browser
http://nciterms.nci.nih.gov

<table>
<thead>
<tr>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI Metathesaurus: National Cancer Institute Metathesaurus (2012 03)</td>
</tr>
<tr>
<td>NCI Thesaurus: National Cancer Institute Thesaurus (12.05d)</td>
</tr>
<tr>
<td>ChEBI: Chemical Entities of Biological Interest (March 2012)</td>
</tr>
<tr>
<td>CTCAE: Common Terminology Criteria for Adverse Events (4.03)</td>
</tr>
<tr>
<td>GO: Gene Ontology (July 2012)</td>
</tr>
<tr>
<td>HUGO Gene Nomenclature Committee (July 2012)</td>
</tr>
<tr>
<td>HL7: HL7 Vocabulary (V3 02-08)</td>
</tr>
<tr>
<td>ICD-10: International Classification of Diseases, Tenth Revision (2012)</td>
</tr>
<tr>
<td>LOINC: Logical Observation Identifiers Names and Codes (234)</td>
</tr>
<tr>
<td>MA: Anatomical Dictionary for the Adult Mouse (March 2012)</td>
</tr>
<tr>
<td>MedDRA: Medical Dictionary for Regulatory Activities Terminology (15.0)</td>
</tr>
<tr>
<td>MGED Ontology: Microarray Gene Expression Data Ontology (1.3.1)</td>
</tr>
<tr>
<td>NPO: Nanoparticle Ontology (1.0_Doc_08_2011)</td>
</tr>
<tr>
<td>OBI: Ontology for Biomedical Investigations (OBI San Diego 2011 Release)</td>
</tr>
<tr>
<td>PDO: Physician Data Query (2012-03-09)</td>
</tr>
<tr>
<td>RadiLex: Radiology Lexicon (3.33)</td>
</tr>
<tr>
<td>SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms (2011_0731)</td>
</tr>
<tr>
<td>UMLS SemNet: UMLS Semantic Network (3.2)</td>
</tr>
<tr>
<td>Zebrafish: Zebrafish Model Organism Database (June 18, 2012)</td>
</tr>
</tbody>
</table>
Terminology best practices have emerged over many decades and can make a big difference to quality and utility. Ontological features are increasingly important.

- Concept-based (one code for one meaning; may have multiple terms or synonyms).
- Human (text) and computer-readable definitions.
- Logical parent-child hierarchies, with poly-hierarchy
- Unique, permanent, meaningless codes or identifiers
- Concept history mechanism to track data over time
- Ontological structure (formal knowledge representation as a set of concepts and the relationships among them) allowing more and better use of data
Concepts: Unambiguous Meaning

Concept Codes
Definitions
Semantic Types
Source specific terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Source</th>
<th>Type</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>NCI-GLOSS</td>
<td>PT</td>
<td>CDR00000659802</td>
</tr>
<tr>
<td>mol</td>
<td>HL7</td>
<td>AB</td>
<td>12650</td>
</tr>
<tr>
<td>mol(s)</td>
<td>ICH</td>
<td>PT</td>
<td>12656</td>
</tr>
<tr>
<td>mole (chemical)</td>
<td>NCI-GLOSS</td>
<td>SY</td>
<td>CDR00000659802</td>
</tr>
</tbody>
</table>

7 sources
Preferred Name: Gastric Mucosa-Associated Lymphoid Tissue Lymphoma
Code: C5266
Semantic Type: Neoplastic Process
Parent Concepts: Extranodal Marginal Zone B-Cell Lymphoma of Mucosa-Associated Lymphoid Tissue
Gastric Non-Hodgkin's Lymphoma

Synonyms & Abbreviations:
Gastric MALT Lymphoma
Gastric MALToma
MALT Lymphoma of the Stomach
MALToma of the Stomach
Primary Gastric MALT Lymphoma
Primary Gastric B-Cell MALT Lymphoma
Primary MALT Lymphoma of the Stomach

Definition: A low grade, indolent B-cell lymphoma, usually associated with Helicobacter Pylori infection. Morphologically it is characterized by a dense mucosal atypical lymphocytic (centrocyte-like cell) infiltrate with often prominent lymphoepithelial lesions and plasmacytic differentiation. Approximately 40% of gastric MALT lymphomas carry the t(11;18)(q21;q21). Such cases are resistant to Helicobacter Pylori therapy.
Best Practices: NCIt Example (Part 2 of 2)

Role Relationships (subset) for C5266 Gastric Mucosa-Associated Lymphoid Tissue Lymphoma:

**Molecular abnormalities:**
- Disease_May_Have_Cytogenetic_Abnormality: Trisomy 3
- Disease_May_Have_Cytogenetic_Abnormality: Trisomy 18
- **Role group 1:**
  - Disease_May_Have_Cytogenetic_Abnormality: t(11;18)(q21;q21)
  - Disease_May_Have_Molecular_Abnormality: AP12-MLT Fusion Protein Expression

**Histogenesis:**
- Disease_Has_Normal_Cell_Origin: Post-Germinal Center Marginal Zone B-Lymphocyte

**Pathology:**
- Disease_Has_Abnormal_Cell: Centrocyte-Like Cell
- Disease_May_Have_Abnormal_Cell: Neoplastic Monocytoid B-Lymphocyte
- Disease_May_Have_Abnormal_Cell: Neoplastic Plasma Cell
- Disease_May_Have_Finding: Lymphoepithelial Lesion

**Anatomy:**
- Disease_Has_Primary_Anatomic_Site: Stomach
- Disease_Has_Normal_Tissue_Origin: Gut Associated Lymphoid Tissue

**Clinical information:**
- Disease_Has_Finding: Primary Lesion
- Disease_May_Have_Finding: Indolent Clinical Course
- Disease_May_Have_Associated_Disease: Hepatitis C
Research focuses on areas of rapid change and new insights, where terminologies have to be highly responsive to be useful.

- e.g. Drugs/Therapeutic agents: a major focus, linked to non-investigational use and coding, requiring joint efforts to share work and build standards with major partners.
  - 24-hour turnaround required for some drug information.
  - Must reflect new classifications, data, warnings, etc.

➤ Clinical trials and research are increasingly international and cross-cutting. Terminology that can be freely shared globally is important, and a major contributor to wider standardization.

➤ Federated development of the range of resources required to fully represent the research enterprise is both an effective, and an essential, success strategy.
Terminology Development and Maintenance

Terminology products need consistent, reliable maintenance, and a stable entity for production and distribution to end users.

- What’s out there and is it extensible on your timelines?
- Consider compatibility issues with regulatory products such as MedDRA.
- Can they accommodate rapid development schedules and quick turnaround times?
- Are there business rules for change control and versioning?
- Do license restrictions limit reuse of content? Who has rights to content once conveyed?
- Are robust, multiple mechanisms provided for contact?
- What user services are provided such as sub-setting?
- Are readily consumable access methods and output formats provided?
Example: NCI Thesaurus (NCIt) Profile

- Published Monthly
- Public domain, open content license
- 100,000 “Concepts” hierarchically organized
- Description-logic based
- Text definitions
- Concept History
- Available on-line and by download (OWL, flat files and LexGrid XML)
- Accessed through LexEVS 5.1 and LexEVS 6.0 APIs. LexEVS 5.1 API is scheduled for deprecation Winter 2013
- Broad coverage of cancer and clinical research domains including prevention and treatment trials
Managing Diversity Through Collaboration

Issue: How to meet core and legacy requirements while responding to new science and biomedical information.

- Many widely used mainstay standards are also the least compliant with best practices, and slowest to respond to new research requirements, while
- More responsive, pure, best-practice efforts often concentrate on specialized domains and have restricted user bases.

Collaborative strategy for convergence and harmonization.

- We work with old-line standards such as ICD(-11) and MedDRA to make them better and harmonize where possible.
- We work with focused groups in specialized areas of interest such as Drugs, Animal Models and Therapeutic Areas.
- Many of our most successful efforts are mid-sized, gathering a critical mass of interested partners with cross-cutting representation in the field of interest.
Collaboration and Integration

- Harmonized Standard Terminology for Therapeutic Agents, Clinical Research, Adverse Events, Clinical Care:
  - Across NCI and multiple NIH ICs
  - FDA (>16K concepts) and other federal agencies
  - Drug Information Resources [NCI Drug Dictionary, Federal Medication Terminologies (FMT), National Council for Prescription Drug Programs (NCPDP) ePrescribing Standards, FDA Structured Product Labeling (SPL), Regulated Product Submissions (RPS), other]
  - CDISC (>10K Concepts) SDTM, CDASH, SEND (Animal Models), Therapeutic Areas (Parkinson’s, Tuberculosis, etc.)
  - Consortia such as Duke Clinical Research Institute, ACC, NHLBI and CDISC Cardiology Standards; NICHD and Neonatal Research Network Terminology (NRNT) for Pediatric and Neonatal Standards
  - Biomarker Terminology and other Basic Science
  - Safety reporting including CTCAE, Device Safety, ICSR (FDA), other
  - Metadata and Domain Model Annotation such as for BRIDG
NCIt Browser by subset
http://ncit.nci.nih.gov/ncitbrowser/pages/subset.jsf
NCI Term Browser Value Sets

http://ncit.nci.nih.gov/ncitbrowser/ajax?action=create_src_vs_tree
Unified Terminology Services

LexEVS APIs  NCI Term Browser  EVS Files

Biomedical Community

LexEVS

NCPDP

CTC
Adverse Events

National Institutes of Health

FDA

SCCM

CDC

LOINC

ISO

SNOMED CT

UCUM

World Health Organization

RxNorm

the Gene Ontology
Margaret Haber
Associate Director
Enterprise Vocabulary Services
National Cancer Institute
mhaber@mail.nih.gov