W3C Semantic Web
Health Care and Life Sciences Interest Group
BioRDF Teleconference
September 22, 2008

The UMLS and the Semantic Web

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Outline

- The UMLS (in a nutshell)
  - Lexical resources
  - Metathesaurus
  - Semantic Network
- Why is the UMLS relevant to the Semantic Web?
- Issues and challenges

Unified Medical Language System (UMLS)

UMLS: 3 components

- SPECIALIST Lexicon
  - 200,000 lexical items
  - Part of speech and variant information
- Metathesaurus
  - 5M names from over 100 terminologies
  - 1M concepts
  - 16M relations
- Semantic Network
  - 135 high-level categories
  - 7000 relations among them

Lexical resources

Terminological resources

Ontological resources

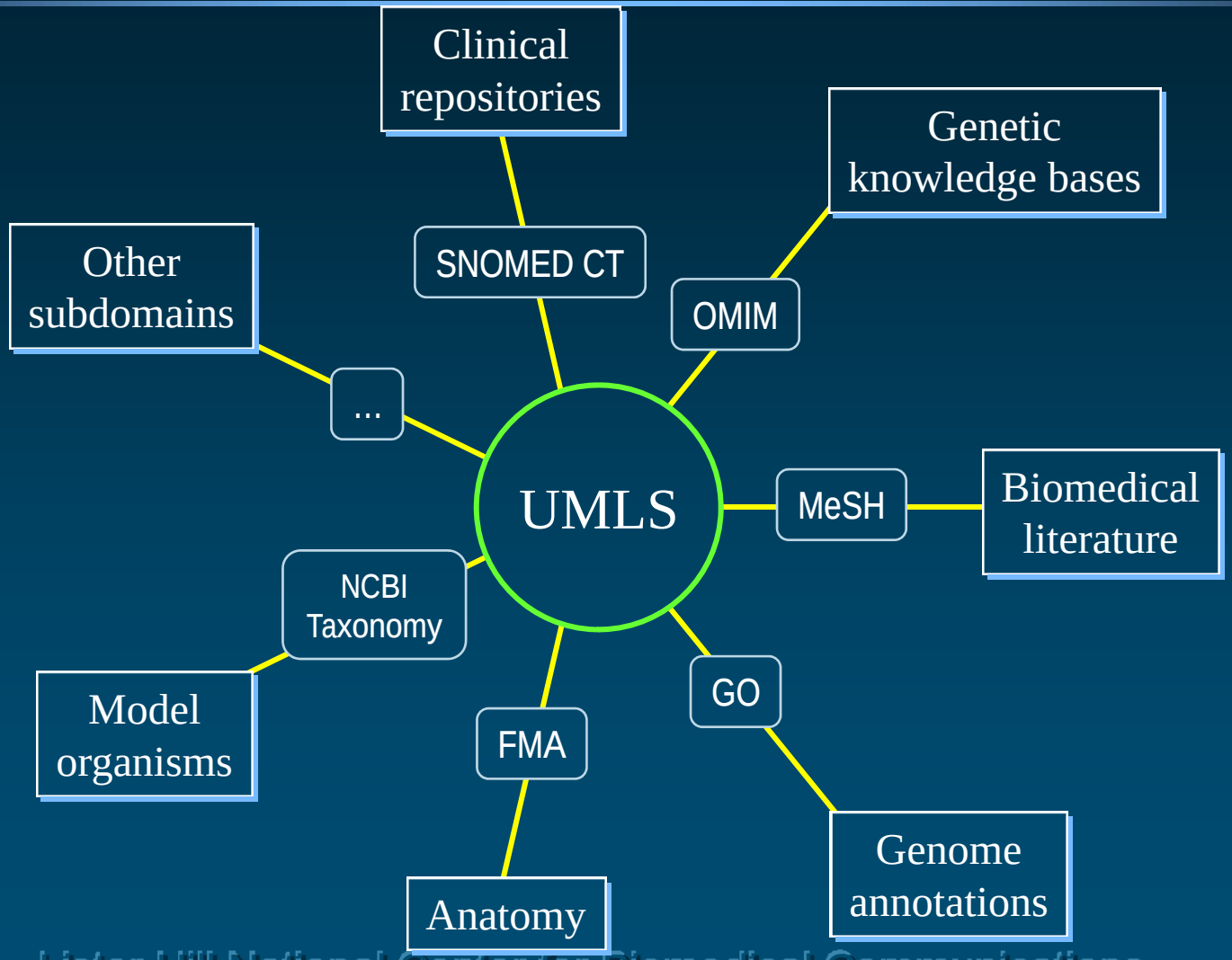
UMLS Characteristics (1)

- Current version: 2008AA (2-3 annual releases)
- Type: Terminology integration system
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (intramural)
- Availability
  - Publicly available: Yes* (cost-free license required)
  - Repositories: UMLS
- URL: http://umlsks.nlm.nih.gov/

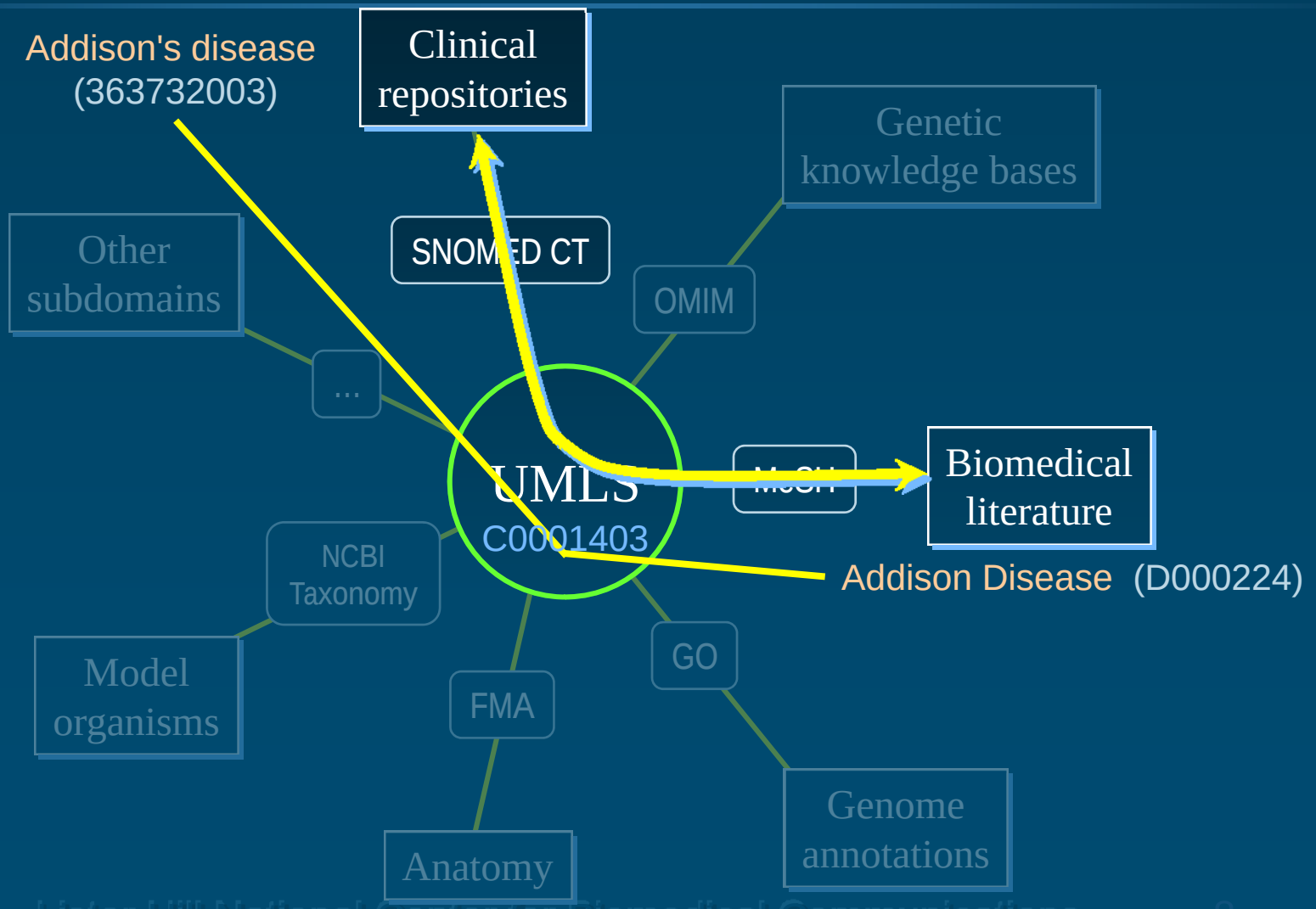
UMLS Characteristics (2)

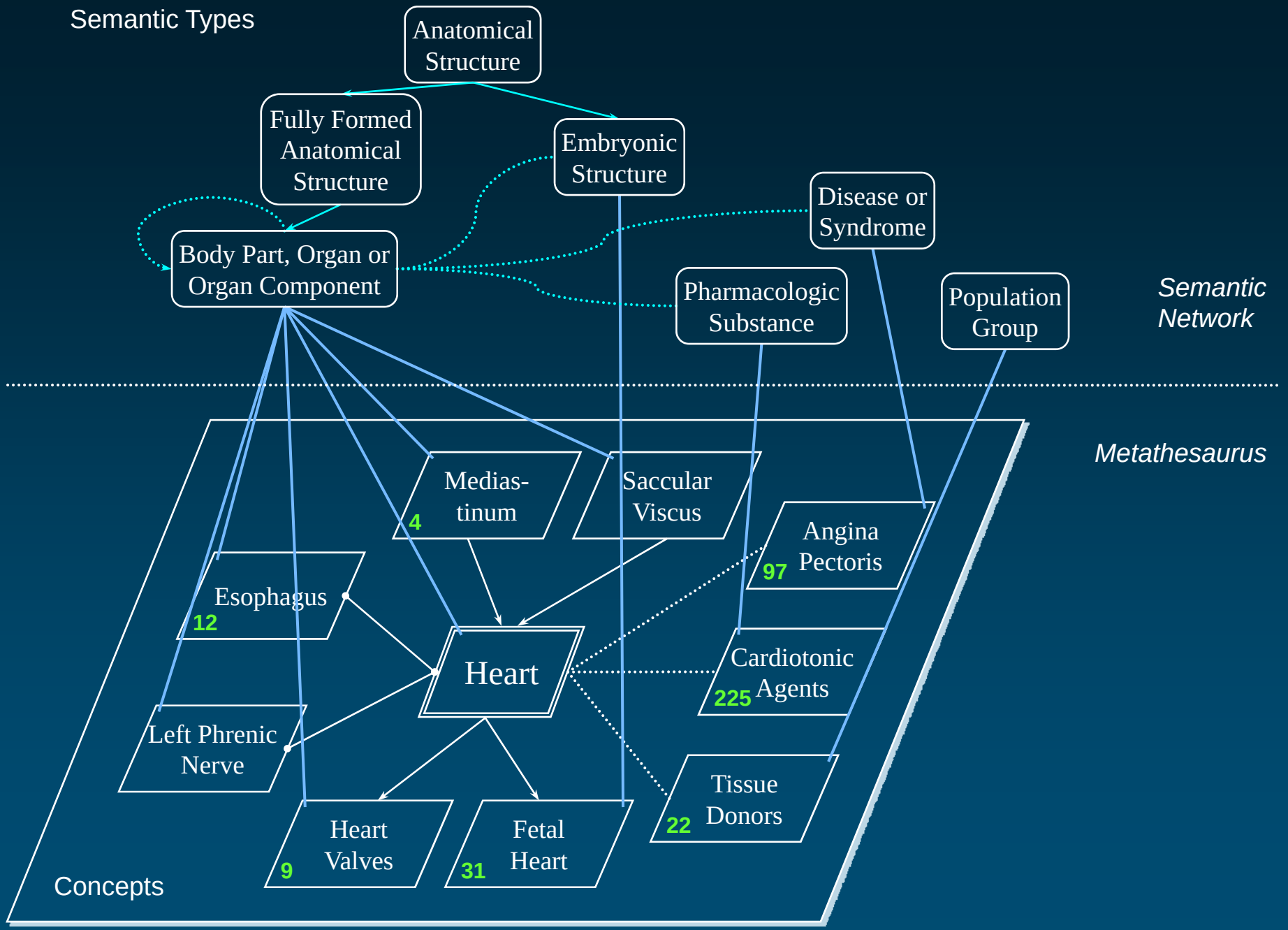
- Number of
  - Concepts: 1.5M (2008AA)
  - Terms: ~6M
- Major organizing principles (Metathesaurus):
  - Concept orientation
  - Source transparency
  - Multi-lingual through translation
- Formalism: Proprietary format (RRF)

UMLS Integrating subdomains



Trans-namespace integration




Semantic Types
Anatomical Structure
Fully Formed Anatomical Structure
Embryonic Structure
Disease or Syndrome
Body Part, Organ or Organ Component
Pharmacologic Substance
Population Group
Semantic Network
Metathesaurus
Medias-tinum
4
Saccular Viscus
Angina Pectoris
97
Esophagus
12
Heart
Cardiotonic Agents
225
Left Phrenic Nerve
Tissue Donors
22
Heart Valves
9
Fetal Heart
31
Concepts

Why is the UMLS relevant to the Semantic Web?

Relevance to the SW Metathesaurus

- Terminology integration system
  - Trans-namespace integration
  - Integration beyond shared identifiers
- Repository of biomedical terminologies/ontologies
- Many UMLS vocabularies used for the annotation of datasets (including clinical records)

Relevance to the SW Metathesaurus

- Broad coverage of biomedicine
- Large user base
- Tooling available
  - E.g, visualization, named entity recognition, etc.

Relevance to the SW Semantic Network

- Top-level ontology of the biomedical domain
- Broad biomedical categories
- Helps partition biomedical concepts
- Semantic relations

Issues and Challenges

Issues and challenges

- Availability
  - Mandatory license agreement
- Discoverability
  - No metadata
- Formalism
  - No easy conversion to SKOS/RDF(S)/OWL
- Identifiers

- Steep learning curve

Availability

- Some source vocabularies have intellectual property restrictions
  - E.g., most drug vocabularies
  - Complex agreement for SNOMED CT: available at no cost for member countries of the IHTSDO
- Mandatory license agreement
  - No cost for research
  - May require negotiation with the vocabulary developer for production applications
- MetamorphoSys helps extract selected sources from the UMLS

Discoverability

- Discoverability of individual concepts
  - UMLSKS web services
  - Search all UMLS source vocabularies at the same time
  - Named entity recognition/normalization (e.g., MetaMap)
- Discoverability of terminologies/ontologies
  - No comprehensive registries
  - No rich registries
    - With rich metadata supporting the discoverability of terminologies/ontologies

Formalism

- UMLS: Proprietary format
  - Rich Release Format (RRF)
  - All terminologies/ontologies represented in the same format
- No easy conversion to SKOS/RDF(S)/OWL
  - Underspecified semantics
    - Child/parent ≠ subClassOf
  - Complex semantics
    - Descriptors / concepts / terms
  - Rich attribute set

Identifiers for biomedical entities

- What is identified?
  - Entity vs. resource about the entity
- Which identifier to pick?
  - E.g., Addison's disease
    - 363732003 (SNOMED CT)
    - D000224 (MeSH)
    - C0001403 (UMLS Metathesaurus)
- Which format?
  - URI vs. LSID
- Which authoritative source for minting URIs?
  - Ontology developers vs. (e.g.) Bio2RDF

Steep learning curve

- Large resource
  - 1.5M concepts
  - 6M terms
  - Over 20M relations
- Complex structure
  - Metathesaurus
  - Semantic Network
- Rich set of attributes
- Rich set of relations
  - Terminological
  - Semantic
  - Statistical
  - Mapping
- Multiple languages
- Complex domain

Conclusions

Conclusions

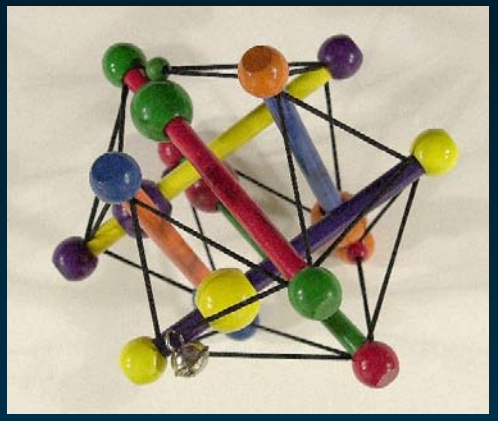
- UMLS as a terminology integration system
  - Helps bridge across namespaces
  - Helps integrate information sources
    - Beyond shared identifiers
- UMLS as a repository of terminologies/ontologies
  - Single source, single format for 143 vocabularies
- Issues with availability, discoverability and formalism
- Identifiers for biomedical entities

References

- UMLS
  umlsinfo.nlm.nih.gov
- UMLS browsers
  (free, but UMLS license required)
  - Knowledge Source Server: **umlsks.nlm.nih.gov**
  - Semantic Navigator:
    http://mor.nlm.nih.gov/perl/semnav.pl
  - RRF browser
    (standalone application distributed with the UMLS)

References

- Recent overviews
  - Bodenreider O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*; D267-D270.
  - Bodenreider O. From terminology integration to information integration: Unified Medical Language System (UMLS). BioRDF Teleconference, W3C Semantic Web Health Care and Life Sciences Interest Group, June 5, 2006. http://mor.nlm.nih.gov/pubs/pres/060605-BioRDF.pdf



Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA