Coming to Terms with the Biomedical Tower of Babel

Implications for the design of a biomedical knowledge network

Alexa T. McCray

Center for Biomedical Informatics
Harvard Medical School

Board on Research Data and Information
National Academy of Sciences
February 26, 2013
“No aspect of human life has escaped the impact of the Information Age, and perhaps in no area of life is information more critical than in health and medicine.”

U.S. National Academy of Engineering
Grand Challenges 2008

http://www.engineeringchallenges.org
The Promise of Genomic Medicine

1990-2003
Human Genome Project

2004-2010

2011-2020

Beyond 2020

Green ED. Nature 2011; 470(10):203-13
The Promise of Genomic Medicine

Understanding the structure of genomes

1990-2003
Human Genome Project

2004-2010

2011-2020

Beyond 2020

The Promise of Genomic Medicine

Understanding the biology of genomes
Understanding the biology of disease

1990-2003
Human Genome Project

2004-2010

2011-2020

Beyond 2020

The Promise of Genomic Medicine

Advancing the science of medicine

1990-2003
Human Genome Project

2004-2010

2011-2020

Beyond 2020

Green ED. Nature 2011; 470(10):203-13
The Promise of Genomic Medicine

1990-2003
Human Genome Project

2004-2010

2011-2020

Beyond 2020

Improving the effectiveness of healthcare

Green ED. Nature 2011; 470(10):203-13
An Inflection Point

“Biomedical research and the practice of medicine, separately and together are reaching an inflection point: the capacity for description and for collecting data, is expanding dramatically, but the efficiency of compiling, organizing manipulating these data – and extracting true understanding of fundamental biological processes, and insights into human health and disease, from them – has not kept pace.”

Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease
Precision Medicine

Individual-Centric

Toward Precision Medicine
National Research Council, 2011
Precision Medicine

Toward Precision Medicine
National Research Council, 2011
**Precision Medicine**

![Diagram of Precision Medicine](image)

**Individual-Centric**

*Toward Precision Medicine*
National Research Council, 2011
Precision Medicine

Individual-Centric

Toward Precision Medicine
National Research Council, 2011
A New Taxonomy of Disease

“Could it be that something as fundamental as our current system for classifying diseases is actually inhibiting progress?”

Toward Precision Medicine. National Research Council, 2011:10
Wide Range of Data Sources Needed

- Laboratory Data
- Clinical Systems
- Biosensor Data
- Social Networking Sites

All have their own

- Syntax
- Semantics

- Some use standard terminologies/ontologies
- Some use home-grown terminologies

Public Databases

- NCBi
  - Literature Databases
  - Entrez Databases
  - Nucleotide Databases
  - Genome-Specific Resources
  - Tools for Data Mining
  - Tools for Sequence Analysis
  - Tools for 3D Structure Display and Similarity Searching
  - Maps
  - Collaborative Cancer Research
  - FTP Download Sites
  - Resource Statistics

Harvard Medical School
Center for Biomedical Informatics
Terminology Use

Coding clinical data
Indexing and retrieving the literature
Annotating genomic data
Statistical reporting, epidemiologic studies
Outcomes measurement
Public health surveillance
Cost analysis
Information exchange and data integration
Data mining, aggregation
Natural language processing
...

Common Terminologies

Varying scope, coverage, rigor, and update schedules
How do we come to terms with this Biomedical Tower of Babel?

- Recognize the value of curating data with standard terminologies
- Recognize that a variety of communities of practice exist
  - Encourage consistency within those communities of practice
- Discourage development of “redundant” terminologies/ontologies
- Map terminologies to each other for maximum semantic interoperability
- Develop robust NLP tools that take advantage of existing terminologies

Unified Medical Language System (UMLS) is one large-scale effort that has made progress toward facilitating semantic interoperability of biomedical data
**UMLS**

- Integrates > 100 existing terminologies/ontologies
- Includes a higher level ontology
- Includes natural language processing tools and resources

**UMLS Knowledge Sources**
- Metathesaurus
- Semantic Network
- SPECIALIST NLP Tools
## UMLS Metathesaurus Mapping Example

<table>
<thead>
<tr>
<th>Medical Term</th>
<th>CUI</th>
<th>LUI</th>
<th>SUI</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renal Cell Carcinoma</td>
<td>C0007134</td>
<td>L0007134</td>
<td>S0399999</td>
<td>MeSH</td>
</tr>
<tr>
<td>Renal Cell Carcinomas</td>
<td>C0007134</td>
<td>L0007134</td>
<td>S0081526</td>
<td>MeSH</td>
</tr>
<tr>
<td>Grawitz Tumor</td>
<td>C0007134</td>
<td>L0018219</td>
<td>S0375417</td>
<td>SNOMED-CT</td>
</tr>
<tr>
<td>Hypernephroma</td>
<td>C0007134</td>
<td>L0020489</td>
<td>S0050424</td>
<td>MedDRA</td>
</tr>
<tr>
<td>Nephroid Carcinoma</td>
<td>C0007134</td>
<td>L0027710</td>
<td>S0065974</td>
<td>MeSH</td>
</tr>
<tr>
<td>Adenocarcinoma of kidney</td>
<td>C0007134</td>
<td>L0493923</td>
<td>S4147170</td>
<td>NCI Thesaurus</td>
</tr>
<tr>
<td>Renal Carcinoma</td>
<td>C0007134</td>
<td>L0161908</td>
<td>S1629563</td>
<td>OMIM</td>
</tr>
</tbody>
</table>
UMLS Two-Level Structure
Neoplastic Process

**Definition**
A new and abnormal growth of tissue in which the growth is uncontrolled and progressive. Growth may be malignant or benign.

**Properties**
- Unique Identifier: T191
- Tree Number: B2.2.1.2.1.2
- Usage Note: All neoplasms are assigned to this type. Do not also assign a type from the 'Anatomical Abnormality' hierarchy.

**Parents**
- Disease or Syndrome

**Relations**
- Neoplastic Process isa Disease or Syndrome (DNI)
- Inherited Relations
- Inverse Inherited Relations
Terminologies in Use

ClinicalTrials.gov

Harvard Catalyst

Correlating Phenotype & Genotype in Autism Research
**ClinicalTrials.gov**

A service of the U.S. National Institutes of Health

---

**Find Studies** > **Search Results**

75 studies found for: Lou Gehrig's disease

Modify this search | How to Use Search Results

---

### Show Display Options

- **List**
- **By Topic**
- **Open Studies**

---

<table>
<thead>
<tr>
<th>Rank</th>
<th>Status</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Recruiting</td>
<td><strong>Phase II/III Randomized, Placebo-controlled Trial of Arimocimol in SOD1 Positive Familial Amyotrophic Lateral Sclerosis</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Conditions:</strong> Amyotrophic Lateral Sclerosis; Acquired Amyotrophic Lateral Sclerosis; Amyotrophic Lateral Sclerosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Intervention:</strong> Drug: Arimocimol</td>
</tr>
<tr>
<td>2</td>
<td>Recruiting</td>
<td><strong>Genetics of ALS: Identification of Genes With Roles in Familial and Sporadic Amyotrophic Lateral Sclerosis (ALS) and Amyotrophic Lateral Sclerosis (ALS) With Frontotemporal Dementia</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Conditions:</strong> Amyotrophic Lateral Sclerosis; Familial Amyotrophic Lateral Sclerosis; Amyotrophic Lateral Sclerosis With Frontotemporal Dementia; Lou Gehrig's Disease; Motor Neuron Disease; Primary Lateral Sclerosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Intervention:</strong> Other: Genetic study of ALS families</td>
</tr>
<tr>
<td>3</td>
<td>Recruiting</td>
<td><strong>The National Amyotrophic Lateral Sclerosis Registry</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Condition:</strong> Amyotrophic Lateral Sclerosis</td>
</tr>
</tbody>
</table>
Recognized Terms and Synonyms:

lou gehrigs disease: 214 studies
als1
amyotrophic lateral sclerosis
autosomal dominant als
bulbar motor neuron disease
charcot disease
familial amyotrophic lateral sclerosis
gehrig disease
lou gehrig disease
motor neuron disease, amyotrophic lateral sclerosis
Apolipoproteins E are a class of protein components which can be found in several lipoproteins including high-density lipoproteins; very-low-density lipoproteins; and chylomicrons. Synthesized in most organs, Apo E is important in the global transport of lipids and cholesterol throughout the body. Apo E is also a ligand for LDL receptors (receptors) in LDL-mediated binding, internalization, and catabolism of lipoprotein particles in cells. There are several allelic isoforms (such as E2, E3, and E4). Deficiency or defects in Apo E are causes of hyperlipoproteinemia type III.

**Apolipoproteins E**

"Apolipoproteins E" is a descriptor in the National Library of Medicine's controlled vocabulary thesaurus, MeSH (Medical Subject Headings). Descriptors are...

HYMAN, BRADLEY T

Neurology. Massachusetts General Hospital.
Matching Topic: Apolipoproteins E (42 publications).

Metabolism of very-low-density lipoprotein and low-density... Medscape. PubMed. Details.

Arterioscler Thromb Vasc Biol. Feb 1, 2010. MENDIVIL CO, ZHENG C, FURTADO JD, LEL J, SACKS FM.
Matching Topic: Apolipoproteins E.

Apolipoprotein E (APOE) genotype has dissociable effects on memory and attention-executive network function in Alzheimer's disease. Proc Natl Acad Sci U S A.
Metabolism of very low-density lipoprotein and low-density... Medvane PubMed details 

Atheroscler Thromb Vasc Biol. Feb 1, 2010. MENDIVIL CO, ZHENG C, FURTADO JD, LEL J, SACKS FM
Matching Topic: Apolipoproteins E

The protective role of adiponectin in pulmonary vascular disease. Medvane PubMed details 

Matching Topic: Apolipoproteins E

Hybrid in vivo FMT-CT imaging of protease activity in atheroscl... Medvane PubMed details 

Matching Topic: Apolipoproteins E

Expression of neuronal nitric oxide synthase splice variants... Medvane PubMed details 

Matching Topic: Apolipoproteins E

A genome-wide genetic screen for host factors required for... Medvane PubMed details 

Matching Topic: Apolipoproteins E

Vascular effects of a low-carbohydrate high-protein diet. Medvane PubMed details 

Proc Natl Acad Sci U S A. Sep 8, 2009. FOO SY, HELLER ER, WYKRYKOWSKA JJ, SULLIVAN CJ, MANNING-TOBIN JJ, MOORE KJ, GERSZTEN RE, ROSENZWEIG A
Matching Topic: Apolipoproteins E
Correlating Phenotype & Genotype

Phenotypic and Genetic Factors in Autism Spectrum Disorders

- ~ 500 families participated
- Data collected
  - Phenotypic data from 25 behavioral instruments
  - Biological samples
- Developed autism-specific ontology
  - Aligned and integrated terminology represented in phenotypic instruments
  - Linked to UMLS whenever possible
  - Ontology will be openly available
Autism Ontology

Scope

- ~5,000 questions from 25 instruments mapped into ~300 concepts
- Three high-level groupings
  - Personal Traits
  - Social Competence
  - Medical History

Use

- Concept-based search of Autism Consortium phenotype data
  - Genotypic data correlated with phenotypic characteristics
- Query, aggregate, and integrate data for hypothesis-driven research
Concluding Remarks

Precision Medicine is a Grand Challenge that is worth pursuing.

Its realization will involve addressing many social, legal, ethical, economic, and technical issues.

If we put our minds to it, the technical issues can be among the first to be solved, and may even show the way to the resolution of some of the even thornier issues.

Semantic interoperability is one technical issue that can be solved.