There are three UMLS Knowledge Sources: the Metathesaurus®, the Semantic Network, and the SPECIALIST Lexicon. They are distributed with flexible lexical tools and the MetamorphoSys install and customization program.

http://www.nlm.nih.gov/research/umls/

2.5 International Nomenclature of Diseases (IND)

3 Data models

A consensus data model will support the effective use and reuse of ICD data for multiple purposes. The definition of disease and of external causes of diseases has been approached in various ways, and several initiatives have created their own model accommodating a specific use. The differences result in incompatible conceptualization and hamper development of specialty and business overarching use of information.

4 Ontological standards

The linkage to and use of ontological concepts and practices will enable ICD to be computable. This is expected to support the global exchange of semantic data. Ultimate interoperability would mimic or exceed that of the financial services industry. Ontologies in the field of information technology aim at representation of specific knowledge. Evolution in this field is fast. A set of standards for representation of ontologies, and tools for handling ontologies have been developed.

4.1 OWL

The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies, and is endorsed by the World Wide Web Consortium. This family of languages is based on two semantics: OWL-DL and OWL-Lite semantics are based on Description Logics which have attractive and well-understood computational properties, while OWL-Full uses a novel semantic model intended to provide compatibility with RDF Schema. OWL ontologies are most commonly serialized using RDF/XML syntax. OWL is considered one of the fundamental technologies underpinning the Semantic Web, and has attracted both academic and commercial interest.

4.2 RDF

Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata model but which has come to be used as a general method of modelling information, through a variety of syntax formats.

The RDF metadata model is based upon the idea of making statements about resources in the form of subject-predicate-object expressions, called triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

5 Technological standards

5.1 SQL3

SQL3 development includes temporal relationships that could be used in the context of ICD revision tracking and management.

ANSI (X3H2) and ISO (ISO/IEC JTC1/SC21/WG3) SQL standardization committees have for some time been adding features to the SQL specification to support object-oriented data management. The current version of SQL in progress including these extensions is often referred to as "SQL3" [ISO96a, b].SQL3 object facilities primarily involve extensions to SQL's type facilities; however, extensions to SQL table facilities can also be considered relevant. Additional facilities include control structures to make SQL a computationally complete language for creating, managing, and querying persistent object-like data structures. The added facilities are intended to be upward compatible with the current SQL92 standard (SQL92).This and other sections of the Features Matrix describing SQL3 concentrate primarily on the SQL3 extensions relevant to object modelling. However, numerous other enhancements have been made in SQL as well [Mat96]. In addition, it should be noted that SQL3 continues to undergo development, and thus the description of SQL3 in this Features Matrix does not necessarily represent the final, approved language specifications.

http://etrij.etri.re.kr/Cyber/servlet/GetFile?fileid=SPF-1041761245678

6 Reference projects and initiatives

The projects and initiatives below will inform the development of ICD from tooling to content to structure.

6.1 The National Center for Biomedical Ontology

The goal of the Center is to support biomedical researchers in their knowledge-intensive work, by providing online tools and a Web portal enabling them to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice. A major focus of the work involves the use of biomedical ontologies to aid in the management and analysis of data derived from complex experiments.

The Center is organized into six core components:

Core 1: Computer science research

Core 2: Bioinformatics research

Core 3: Driving biological projects and external research collaborations

Core 4: Infrastructure

Core 5: Education

Core 6: Dissemination

The Center is a US National center, assembling the expertise of leading investigators from across the country.

6.2 Organization of the Center

The Core 1 computer-science research and the Core 2 bioinformatics research involve the participation of Stanford University, Lawrence Berkeley National Laboratory, Mayo Clinic, University

of Victoria, and University at Buffalo. Two Driving Biological Projects involve investigation of model-organism databases (FlyBase and ZFIN), while the third involves analysis of clinical-trial data stored in TrialBank.

The computer-science research in Core 1 delivers tools for accessing and unifying ontologies, and Core 2 concentrates on creating tools for using these ontologies to annotate large biomedical data sets, enabling data-set analysis, and integration. These tools enable the driving biological projects in Core 3. There is a direct flow of tools and technologies from Core 1 to Core 2 to Core 3, while the projects in Core 3 motivate the Center's research activities at all levels.

The Center achieves its objectives by advancing standards of good practice, by creating tools and theories that support a wide range of driving biological projects and collaborative research activities, and by training computational biologists, specialists in informatics, and computer scientists in the use of ontologies and of the Center's technologies in support of their research.

The National Center for Biomedical Ontology is part of the National Centers for Biomedical Computing supported by the NIH Roadmap. The Center is funded by the National Institutes of Health (NIH) and is part of the network of National Centers for Biomedical Computing.

6.3 Bioportal

BioPortal is a Web application to access the Open Biomedical Ontologies (OBO) library. This library contains a large collection of ontologies in biomedicine spanning many species from Arabidopsis to Homo Sapiens, and many scales, from molecules to whole organism. The ontology content includes the ontologies of the model organism communities, biology, chemistry, anatomy, radiology, and medicine.

BioPortal permits users to browse individual ontologies with three browsing paradigms—text, tree-view, and graph view. In addition, it permits search across all or specific ontologies according to term name or attribute content.

BioPortal will soon be providing a suite of tools for developers to integrate its functionality into their own applications, by providing URIs for all ontology content and Web services for accessing BioPortal functionality. It will also be providing tools to enable the community to comment on ontologies and their contents, to align and map them, as well as novel ways to visualize them and use them in applications. [2008]

http://www.bioontology.org/

6.4 FMA

The Foundational Model of Anatomy ontology (FMA) is an evolving computer-based knowledge source for bioinformatics; it is concerned with the representation of classes and relationships necessary for the symbolic modelling of the structure of the human body in a form that is understandable to humans and is also navigable, parseable, and interpretable by machine-based systems. Specifically, the FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. Its ontological framework can be applied and extended to all other species.

The Foundational Model of Anatomy ontology is one of the information resources integrated in the distributed framework of the Anatomy Information System developed and maintained by the Structural Informatics Group at the University of Washington.

http://sig.biostr.washington.edu/projects/fm/AboutFM.html

6.5 Semantic health

To efficiently implement E-Health to meet the rising needs of mobile citizens, patients and providers, its fragmented interoperability initiatives must come together and coordinate with the increasing need to link clinical data to information from basic biological sciences and evidence of best clinical practice. Considering the need for interoperability at the Member State and cross-border level of the European Union - as expressed in the EU E-Health Action Plan - and for global interoperability - as represented by WHO - it is necessary to embark on a process that will prompt the divergent initiatives to join forces for the benefit of all citizens.

This SemanticHEALTH SSA develops a European and global roadmap for RESEARCH in health-ICT, focusing on semantic interoperability issues of E-Health systems and infrastructures. The roadmap will be based on consensus of the RESEARCH community, and validated by stakeholders, industry, and Member State health authorities. It

identifies key short-term (2-5 years) and medium-term (4-10 years) RESEARCH needs to achieve semantic interoperability of E-health systems (including issues of nomenclatures presently in use, classifications, terminologies, ontologies, EHR and messaging models, public health and secondary uses, and decision support, their relationships, mapping needs, limitations)

analyses unsolved research issues arising in the context of realistic approaches to priority clinical and public health settings (reflecting on models of use, benefits expected, concrete application experience and lessons learned; relevance of open source model)

takes into account the impact of non-technological (health policy, legal, socio-economic) aspects

reflects and integrates results of related FP6 (E-health ERA, I2-Health and other) studies.

The consortium and associated experts represent centres of excellence from four continents and the WHO.

http://www.semantichealth.org/

6.6 Disease ontology

Disease Ontology is a controlled medical vocabulary developed at the Bioinformatics Core Facility in collaboration with the NuGene Project at the Center for Genetic Medicine. It was designed to facilitate the mapping of diseases and associated conditions to particular medical codes such as ICD9CM, SNOMED and others. This mapping is useful if efforts like the NuGene Project because it allows request for particular tissue type requests to be mapped quickly and with high fidelity to a set of ICD9 codes that can then be used to retrieve appropriate samples from the tissue bank. Without such a mapping, clinicians are forced to manually search through ICD9CM coding booklets to find all possible applicable codes matching their request. Given the complex organization of ICD9CM and difficulty of the manual process, codes and therefore tissue samples are often overlooked. In one sample case, an early version of the Disease Ontology doubled concept coverage while reducing the overall misclassification error percentage. Eventually we envision that the Disease Ontology can also be used to associate model organism phenotypes to human disease as well as medical record mining.

Disease Ontology is implemented as a directed acyclic graph (DAG) and utilizes the Unified Medical Language System (UMLS) as its immediate source vocabulary to access medical Ontologies such as ICD9CM. Using this standard, much of the process of updating the ontology can be handled by

UMLS, freeing resources for clinicians to pursue more urgent tasks. For situations where the graph needs to be directly edited, the open source graph editor DAGEdit can be used. DAGEdit can readily manipulate and view the Disease Ontology because it is stored in Open Biomedical Ontologies (OBO) format in order to take advantage DAGEdit and any other OBO standards compliant tools. The screenshot of the Disease Ontology was taken using DAGEdit and show version 3 of Disease Ontology.

As a graph, the Disease Ontology can be thought of as a subset of UMLS. It fills a niche in the medical ontology world as a lightweight ontology offering context-free concept identifiers designed specifically to facilitate mapping to medical billing codes. Other Ontologies such as SNOMED and MESH lack these features.

The previous version of Disease Ontology (v2.1) is based almost entirely on ICD9CM with additional concepts included that are useful for mapping common disease requests. It is a lightweight ontology containing 19136 concept nodes and is currently available for download. The newest version Disease Ontology 3 (revision 21) is based on primarily on freely available UMLS vocabularies (including ICD9) and is currently under development.

http://diseaseontology.sourceforge.net/

http://ontology.buffalo.edu/bfo/BeyondConcepts.pdf

6.7 Knowledgebases

Several groups have compiled structured disease specific knowledge, as in clinical trial databases. However, the models differ across the groups. Aligning and feeding this content into ICD in the context of the revision process would fertilize the revision and allow world wide harmonization of the information model. Comparison of existing structures would allow distilling the most suitable model. Including a genetic axis would allow correlating genotypes and phenotypes of large population samples through data derived from routine data collection. Recruitment of best possible candidates for tests would become easier. Feeding in content from multiple sources will prevent possible bias to the revision.

6.8 caBIG

caBIG™ stands for the cancer Biomedical Informatics Grid™. caBIG™ is an information network enabling all constituencies in the cancer community – researchers, physicians, and patients – to share data and knowledge.

Molecular-based research is generating vast amounts of complex genetic data, and there is a need to integrate this information with separate and distinct clinical data. Furthermore, this research is based on disparate document and data formats, making it difficult to leverage in meaningful ways.

Research advances and clinical improvements face the following challenges:

A huge and growing volume of data – including genomic and proteomic information – that must be collected, analyzed, and made accessible

A multitude of "legacy" or previously developed information systems, most of which cannot be readily shared between institutions. Many of these systems continue to be paper-based, rather than electronic

An absence of common data formats

Few common vocabularies, making it difficult, if not impossible, to interlink diverse research and clinical results

An absence of tools to connect different databases

The mission of caBIG™ is to develop a truly collaborative information network that accelerates the discovery of new approaches for the detection, diagnosis, treatment, and prevention of cancer, ultimately improving patient outcomes.

The goals of caBIG™ are to:

Connect scientists and practitioners through a shareable and interoperable infrastructure

Develop standard rules and a common language to more easily share information

Build or adapt tools for collecting, analyzing, integrating, and disseminating information associated with cancer research and care.

Difficulty in identifying and accessing available resources, such as biospecimens and reagents

An absence of information infrastructure to share data within an institution, or among different institutions

http://cabig.cancer.gov/

7 Software

7.1 Lexgrid

LexGrid provides support for a distributed network of lexical resources such as terminologies and ontologies via standards-based tools, storage formats, and access/update mechanisms.

Currently, there are many terminologies and ontologies in existence. But just about every terminology has its own format, its own set of tools, and its own update mechanisms. The only thing that most of these pieces have in common with each other is their incompatibility. This makes it very hard to use these resources to their full potential. We have designed the Lexical Grid as a way to bridge terminologies and ontologies with a common set of tools, formats, and update mechanisms. The Lexical Grid is:

accessible through a set of common API's

joined through shared indices

online accessible

downloadable

loosely coupled

locally extendable

globally revised

available in web-space on web-time

cross-linked

http://informatics.mayo.edu/LexGrid/

7.2 Protégé

Protégé is a free, open source ontology editor and knowledge-base framework. It supports two main ways of modelling ontologies via the Protégé-Frames and OWL. Protégé is being developed at Stanford University in collaboration with the University of Manchester and is supported by a strong community of developers and academic, government and corporate users, who are using Protégé for knowledge solutions in areas as diverse as biomedicine, intelligence gathering, and corporate modelling.

7.3 ICD Update and Revision Platform

ICD Update and Revision Platform is a web based system that is designed to facilitate communication within expert workgroups. It has been operational since January 2006.

The basic functionality of the software can be summarized as follows:

This is a web application which means that the users access it from a web site without the need for downloading or installing any programs. All of the information collected and shared by the application resides on a central location in the WHO servers.

The application collects update and revision proposals in a structured and organized manner. This is done by asking the user to fill a form in which the user explains the proposal as well as the rationale behind the proposal. In addition to this, he/she can provide links to web pages such as publication references from PubMed or can upload documents that are relevant to the proposal.

The proposals are organized according to the existing ICD-10 structure. So the users may browse through the classification and see what is proposed in each and every part of the classification. They may post their comments on the proposals which are compiled and displayed with the proposals themselves

The application is not only for collecting the proposals but also works as a workflow engine which starts when the proposal is created and ends when it is removed from the system or it is implemented in the ICD. The workflow is in line with the present updating process.

Different users have different levels of authorization in the workflow. For example, standard users can submit proposals but cannot decide whether they are accepted or when they are going to be implemented, etc.

All old versions of proposals due to editing, all deleted and rejected proposals are stored in the system for review and cannot be edited any more.

http://extranet.who.int/icdrevision

7.4 Lexwiki

A collaborative terminology authoring platform based on Semantic MediaWiki which is currently under development at the Mayo Clinic.

8 Recommendations

Linkage between ICD and terminologies, e.g. SNOMED will allow taking advantage of the internal and external linkages of such terminologies, and allowing the terminologies to feed information into established national and international health information systems, as for decision support, or public health. SNOMED CT is a terminology with ontological relationships starting from pathology. It is

pooling different terminologies that are maintained by specialists in the relevant fields. SNOMED is the biggest clinical terminology in the field of health. Linkages between ICD and SNOMED should be part of the revision process. Established linkages to ICD-10 can be a starting point.

Differing information models for "disease" exist. A sett of tools that are relevant to the revision is in use or under development. An additional advisory group for informatics and modelling will align disease models, assess, design, and produce the tooling environment necessary to the revision that is in line with the current technical environment.

---

[i] WHO Nomenclature Regulations, 1967; ICD-10, 1st edition Volume 1 ; p.1241ff

[ii] WHOSIS Health expenditures 2008, data year 2005, ICD implementation in reimbursement and resource allocation, WHO Implementation database

[iii] MDGs