OpenDDX: an open, distributed database of disease-symptom observations

The existence of an open, reliable, global database that associates symptoms with diseases would be a foundational resource for medical researchers and would eventually save millions of lives by facilitating physicians' differential diagnoses and permitting the public to monitor their own health better . While the benefits of symptom-by-disease databases have been recognized for many years, and a number of proprietary diagnostic tools exist, there is no open, multi-stakeholder data resource that could form the basis of an ecosystem of research, public health and commercial applications. However, recent developments in informatics, such as ontologies for diseases and symptoms, semantic web concepts and tools, handheld data collection and dissemination platforms and accessible data mining tools now enable the creation of a distributed data network based on academic data mining and citizen science observations, and we believe the reality of a global diagnostic health resource may be only a few years away.

Key features of the system:

- The data will be generated in two forms:
- Generic statements of the association of a symptom with a disease, with estimates of frequency and diagnostic importance. These data will be relatively easy to generate, from physicians' experience and by mining academic papers. Where possible, citations will be listed for each record.
- Patient-based records of a symptom at a time and place, with patient metadata such as ethnicity, experienced environment and other diseases. These are the core observations from which context-dependent disease-symptom matrices can be generated, but will be more difficult to generate. However, having access to these raw data and applying sophisticated analysis should help alleviate many physicians worries that expert systems that mimic the pattern-recognition abilities of the human brain cannot be built.
- The presentation of the data will explain clearly their nature and their limitations. On its own, this combined dataset will not be a diagnostic system, but *should be of fundamental value* for creating such systems.
- The data are distributed. No single database holds or nor institution controls the data. This makes collaboration more likely, since every institution can host its own data, but does add exciting challenges for data integration.
- The data come from both within academia and from the public. Citizen science efforts are daily proving their value (e.g., in monitoring climate change) and people's' natural interest in their health makes the public a vast resource for meaningful scientific observations. A vital component of this citizen science project will be a careful assessment of which symptoms can be reliably observed by non-physicians, associated with the development of training resources.

Core ontologies

This project is made possible by two OBO ontologies:

- The Disease ontology (DO)
- The Symptom ontology (SYMP)

Another core ontology needed will be a human racial/genetic/ethnicity classification

Applications

Various applications of the data exist:

- As raw data for further scientific research.
- Physician-focused differential diagnostic tools. Especially in circumstances where physician training is limited, having a free, online source of potential differential diagnoses should cause major increases in quality of healthcare. Our test case of a clinic in West Kalimantan, Indonesia will be used to explore this assertion.
- Public/commercial 'home health' applications, focused on end users and patients. Antony's company Senstore already has an interest in such an application.

Successful commercialization of free data resources now has a proven track-record (e.g., companies using Wikipedia, twitter, etc), and we hope that commercial partnerships will help further fund the open data collection process.

Existing potential databases

It may be possible to persuade the holders of one or more existing databases to share their data:

- The Diseases database
- DXplain (Antony has chatted with them)
- DDX-Morph for dermatology
- OMIM, An Online Catalog of Human Genes and Genetic Disorders

Distributed sources

One of the most innovative aspects of this project will be the call for the development of multiple datasets, up to and including personal patient observations of their symptoms. To make this work, some important components will need to be developed, some perhaps for the first time:

- A system of identity authentication: "How can we trust that the person you said they made the annotation is that person?" While work on standards for verifying identities are in progress (see e.g., OIX, OpenID, Digital Identity, Kantara), we might opt for the old, elegant 'Web of Trust' system. RDF documents and statements with a graph URI can be signed using public PGP keys; an ontology for this already exists.

- A system whereby quality assessment that can be assigned to the data providers, so that data can be filtered by reputation. This has to be a social acceptable method, but is vital to be able to separate academic sources from public sources.
- Multiple data-entry applications and databases serving their data as RDF.
- A registry of these data sources, so that all the relevant statements can be found. Using an existing semantic web crawler and indexer is also possible, but it may be more efficient to have a registry.

Data processing

Because of the distributed nature of the data, sophisticated data processing will be needed to collate them, weigh them according to reputation, group patient records by ethnic and geographical class, and integrate these data according to end use. This step will be carried out both at research centers and by companies developing applications based on the data.

Privacy

The core datum is a record of a symptom observed in a patient with a disease. The expression of that symptom may be influenced by environment, genetic background, and existence of other diseases. Thus the more that is known about the patient the better: it is very valuable to be able to tie data together by an identifier for the patient. While it is easy to anonymize the patient in terms of name and serial numbers (e.g. by using identifiers that are hashes of name, email or SSN), the more data that are associated with an individual, the more possible it becomes to associate the anonymous identifier with an actual person. Securing identity will need to be a major concern of this project.

Roadmap

Within 6 months, we should plan to submit two proposals:

- one to NIH for funds to hire assistants to fill in a database for 100 top diseases,
- the other to the Gates foundation more focused on the data management and manipulation challenges and on generating a productive climate for private/public partnership.

Simultaneously, we will:

- Develop the ontology for the data records and the basic tools for data entry (initially based on Protege)
- Enter a small test data set from resources listed by Lynn
- Set up basic database for patient observations at ASRI clinic in West Kalimantan

Other resources

- Cam’s original post
- The FAQ at Diseases Database