NIH/NLM World Wide Web Database Projects

R. P. C. Rodgers <rodgers@nlm.nih.gov>

(The full paper had not been received at the time of printing of the proceedings.)

Extended Abstract

The U.S. National Library of Medicine maintains the world's largest collection dealing with a single scientific or professional topic. In addition to caring for over 4.5 million holdings (including books, journal, reports, manuscripts and audio-visual items), the NLM offers extensive online information services dealing with clinical care, toxicology and environmental health, and basic biomedical research. This is an overview of Internet- and World-Wide Web-based activities currently underway at NLM: these activities are being undertaken by a large number of individuals and groups within the the library's two research components (the Lister Hill National Center for Biomedical Communications and the National Center for Biotechnology Information, or NCBI) as well as various specialized information services.

These activities build upon the library's well-established online services. Chief among these is MEDLINE, one of the first and still one of the most widely consulted bibliographic databases. An outgrowth in the early 1960s of the NLM's printed document, Index Medicus, MEDLINE employs a controlled vocabulary known as Medical Subject Headings (MeSH) to index the contents of over 3,600 journals from the biomedical literature. MEDLINE and various sister databases can be accessed through direct telephonic connection to ELHILL, a mainframe-based facility at NLM. More commonly in recent years, these resources have been accessed via telephone (or, more recently, Internet) via software known as Grateful Med, which exists in versions for the Macintosh and Microsoft Windows.

A second key component underlying NLM's Internet activities is the the Unified Medical Language System (UMLS) Project. Undertaken in the 1980s, this continuing research and development initiative comprises four linked knowledge sources:

o The UMLS Metathesaurus is a thesaurus containing over 125,000 biomedical concepts, which is linked to the contents of over 25 pre-existing defined vocabularies.

o The UMLS semantic network applies high-level descriptions of meaning, known as semantic types, to the contents of the Metathesaurus, as well as supplying a list of semantic relationships that can occur between various semantic types.

o The SPECIALIST lexicon contains computational linguistic information about ordinary English as well as biomedical terminology, and serves as a resource for the SPECIALIST natural language processing system.

o The Information Sources Map (ISM) consists of a database of information resources tagged with descriptive information and indexed using MeSH and elements of the UMLS semantic network.

The World-Wide Web (WWW) plays a prominent role in current NLM innovations. Several Web-based systems will be discussed and demonstrated:

o OnLine Images (OLI): a system for browsing large cataloged image archives, and retrieving selected images. This was applied initially to nearly 60,000 images from the prints and photographs collection of the NLM's History of Medicine Division.

o BankIt: a tool for submitting genetic sequences to NCBI's GenBank database.

o NetCoach (working title): an application that uses the UMLS tools to interact with a user of MEDLINE to help refine searches of the biomedical literature.

o Sourcerer: an application which uses the Information Sources Map and other UMLS tools to identify network-based information resources that are appropriate for answering a specific user-generated query. An associated tool, Apprentice, allows information providers to remotely register new information sources.

World-Wide Web technology provides a number of advantages to NLM in its role an information provider:

o Ease of use for the end-user: once installed properly, a Web client is extremely easy to use.

o Multimedia capabilities: there is much biomedical information in the form of images and

sound.

o Multiprotocol support: this allows the integration of disparate sources and legacy systems into a single access point.

o Platform independence: the availability of multiple Web clients for use with terminals and virtually all commonly used computers has freed NLM from the arduous and expensive task of writing platform-specific clients, and freed it to concentrate on content and information retrieval issues.

Web technology also poses interesting challenges:

o Internet access: many of NLM's users do not yet have high-speed Internet access.

o The rapid evolution of Web technology makes demands on the time of developers who are trying to keep things up-to-date.

o Lack of stateful properties in the underlying protocol: this deficiency has been overcome by use of NCSA's Common Gateway interface (CGI).

o Inconsistent behavior among Web clients: capabilities of various clients vary, requiring the information provider to maintain lists of clients that operate correctly for specific services.

o Resource location: the browsing paradigm followed by the early stages of the Web has proven useless for much biomedical information retrieval, where users often have tightly focussed questions that must be answered swiftly. Navigation aids such as Sourcerer are meant to address this issue.