Transforming Biomedical Informatics and Health Information Access
B.L. Humphreys et al. (Eds.)

doi:10.3233/SHTI210985

Something New and Different: The Unified Medical Language System

Betsy L. HUMPHREYS M.L.S [a, 1] and Mark S. TUTTLE [b]
[a] *U.S. National Library of Medicine (retired)*
[b] *Apelon (retired)*

Abstract. Donald A.B. Lindberg M.D. arrived at the U.S. National Library of Medicine in 1984 and quickly launched the Unified Medical Language System (UMLS) research and development project to help computer understand biomedical meaning and to enable retrieval and integration of information from disparate electronic sources, e.g., patient records, biomedical literature, knowledge bases. This chapter focuses on how Lindberg's thinking, preferred ways of working, and decision-making guided UMLS goals and development and on what made the UMLS markedly "new and different" and ahead of its time.

Keywords. Unified Medical Language System, Donald A.B. Lindberg M.D., U.S. National Library of Medicine

1. Introduction

When Donald A.B. Lindberg M.D. became the Director of the U.S. National Library of Medicine (NLM) in 1984, he strongly believed in the promise of computers to help people provide better patient care [1]. Nevertheless, he had experienced firsthand the difficulties of developing systems that could deliver on that promise. He arrived at NLM with the intention of launching a new informatics research and development effort aimed at reducing those difficulties. His visionary goal was to help computers "understand" biomedical meaning, in essence a "Grand Challenge" that predated use of the term in informatics. Lindberg wanted to enable the retrieval and integration of information from disparate electronic sources, e.g., patient records, biomedical literature, knowledge bases. His plan was in preliminary form, but it had a name: the Unified Medical Language System (UMLS); a specific problem to address: "… the most fundamental barrier to the application of computers in medicine; namely, the lack of a standard language in medicine;" and intended users: developers of computer applications and informatics researchers [2]. This was a novel target user group for him, for NLM, and the field of medical informatics.

Lindberg conceived of the UMLS project in the months between his selection as NLM's next director and his assumption of the position in late August 1984. He foresaw the inevitable exponential growth in the size, diversity, and importance of information sources in digital form. These would be critical in improving health care and biomedical research. He pondered what NLM might do to foster advanced new computer systems

[1] Corresponding Author, Betsy L. Humphreys, 3625 10th Street North, Unit 305, Arlington, VA 22201, USA; Email: betsyhumphreys@verizon.net.

that could retrieve and integrate such disparate sources He was familiar with the difficulties caused by varying vocabularies and codes in different types of medical information. Then unusual in the medical informatics field, Lindberg had experience with diverse sources of data and had built disparate information systems. He had worked with digital lab results and electronic texts; used large, mainframe computers in production systems; and dealt with computational complexities, e.g., attempts to implement expert reasoning. He had served as an advisor to the NLM-funded project that produced the Association of American Medical Colleges (AAMC) report on Integrated Academic Information Management Systems (IAIMS) [3]. He understood that the language problem would become more acute as institutions attempted to integrate networked sources of clinical, administrative, and published research knowledge and then share the results of their efforts for re-use elsewhere.

In preparation for his move to NLM, Lindberg expanded his informal consultations with other medical informatics pioneers to obtain advice on what major new step NLM could take to advance computer applications in medicine. His discussions with Marsden Scott Blois M.D., Ph.D. were particularly influential. Blois published his foundational theory on the requirement for vertical reasoning in medical diagnosis, across multiple levels of information, from the patient as a whole - down to atoms or ions, each with its own vocabulary, in book form in 1984 [4]. Lindberg and Blois had separately used a machine-readable version of the American Medical Association's (AMA) *Current medical information & terminology (CMIT): for the naming and description of diseases and conditions in practice and in areas related to medicine* in pioneering systems that suggested possible diagnoses based on patient presentations [5-8]. Although unarticulated at the time, these activities formed a partial model for the UMLS: digital medical knowledge was provided for use by system developers. Both Lindberg and Blois attempted unsuccessfully to convince the AMA to continue producing updated editions of CMIT, an early effort to name and define all diseases using structured definitions.

By the time he arrived at NLM in August 1984, Lindberg had identified the UMLS as a long-term project that would play to NLM's strengths as a Federal Agency with a track record of technical innovation and development and use of standards. For example, NLM had already had success in building and maintaining the Medical Subjects Headings (MeSH) and large-scale medical information systems used worldwide. Experience had taught him that grant-funded academic institutions and professional associations were not ideally positioned to maintain large terminology resources over time. He intended to enlist both, however, in helping to define, develop, test, and refine what he anticipated as UMLS components. Any resources produced by the UMLS project would be freely available for iterative testing and experimental use by system developers and informatics researchers in the U.S. and other countries. This was another first for NLM, a commitment to "Open Science", again predating that term. It was viewed with concern by some producers of medical terminologies.

Much has been written about the UMLS project and the heavily used resources resulting from it [e.g., 9-13]. This chapter focuses on how Lindberg's thinking, preferred ways of working, and decision-making guided UMLS goals and development and on what made the UMLS markedly "new and different" and ahead of its time.

2. Establishing the UMLS Project

Following his preferred pattern for developing new projects, Lindberg circulated a brief statement about his "fuzzy" UMLS idea soon after his arrival at NLM. He began immediately to solicit input from his senior staff, Board of Regents members, and others encountered at the Library. As he often said, "No one has a lock on good ideas. They come from everywhere." By January 1985, he had established a multidisciplinary NLM UMLS Team. Two months later he asked Congress for additional FY1986 funding for the UMLS project, the first such request during his NLM tenure.

While awaiting the verdict on additional funding, Lindberg consolidated and expanded the NLM UMLS team, which he chaired. Harold M. Schoolman M.D. served as his chief lieutenant during this formative stage. They designated Betsy L. Humphreys M.L.S., as Executive Secretary of the team, which also included Lawrence C. Kingsland III Ph.D. and Peri L. Schuyler M.L.S. Collectively, the initial team had expertise in medicine, chemistry, terminology development, computer science, artificial intelligence, library and information science, standards, database development, production systems, and project and contract management. Lindberg regarded linguistics as an essential missing piece. When the first linguist tried was not a good fit, Lindberg persevered, and Alexa T. McCray Ph.D. joined the team in January 1986. Daniel R. Masys M.D. became a member of the UMLS team when appointed Director of NLM's Lister Hill National Center for Biomedical Communications in June 1986. William T. Hole M.D. was added to the NLM team in January 1989 to play a leading role in UMLS Metathesaurus development and production.

Lindberg was a visionary, but his strategy for advancing toward any large goal was data-driven and incremental. He expected the need for adjustments in response to new knowledge and emerging opportunities, whether in method, technology, or content. During 1985 and early 1986, the NLM UMLS team compared some key biomedical vocabularies and classifications, e.g., NLM's Medical Subject Headings (MeSH), the International Classification of Diseases (ICD), the Systematized Nomenclature of Medicine (SNOMED), to gain a better understanding of the problem the UMLS aimed to solve. In this context, the definition of "key" was *in use* in machine readable biomedical information sources. This early work confirmed that significant differences in the content and structure of terminology systems reflected significant differences in purpose and use. No single vocabulary system was at all likely to meet all anticipated needs.

Based on Schoolman's advice, Lindberg selected the method (Task Order research contracts) for funding the participation of university-based informatics research groups to give NLM more control over evolving major decisions than possible with grant mechanisms. Since NLM would be responsible for long term maintenance of any successful UMLS resources, he needed the final say on their scope and development methods. Lindberg personally enlisted the American Medical Association (AMA) and the College of American Pathologists (CAP) as public allies of the UMLS project, although their corporate views of it would change over time. He also verified that work underway to update the ICD would not reduce the problem the UMLS was intended to address.

Congress added one million dollars to NLM's FY 1986 budget to support the UMLS project. Lindberg allocated an equal amount from NLM's existing research budget. In March 1986, NLM issued a competitive Request For Proposals (RFP) for multiple two-year research and development contracts. The RFP reflected the Library's then-current

thinking about UMLS objectives and strategy, including the probable need to develop at least two new knowledge sources, a Metathesaurus (the word was coined by NLM during the UMLS project) and an Information Sources Map [Endnote 1]. The first two-year contract period was intended to be exploratory, however, and to result in firm decisions about the necessary UMLS components and how to build them, as well as a greater understanding of the context in which they would operate, e.g., medical natural language, existing vocabularies and classifications, machine readable biomedical information sources, and user information needs.

In August 1986, NLM awarded four Task Order research contracts to teams including seven distinguished informatics research groups in five states. Several teams already held NLM-funded informatics training grants. The NLM UMLS team was the eighth group in the sixth state. Humphreys was NLM's technical project officer for the contracts. The list of initial UMLS research participants was a who's who in medical informatics. At least ten were already Fellows of the American College of Medical Informatics elected in its inaugural two years (1984-85), and many who worked on the project would be elected later. The UMLS project was "a distributed national experiment", to use Lindberg's term, and an early U.S. example of a funded "multidisciplinary, multicenter study" in medical informatics research. With no model to follow, NLM and its UMLS contractors proceeded to establish a framework for collaboration, including relatively early use of email via the Internet.[10]

3. Explaining the UMLS Goals and Assumptions

The initial level of confusion about the UMLS goals and general approach may be hard to comprehend today. To those involved in biomedical informatics and data science in the 2020s, the need to retrieve, integrate, and aggregate information and data at scale from disparate machine-readable sources with different terminologies and code sets is obvious. The value of regularly updated multi-purpose resources, whether knowledge sources or programs, to meet this need is apparent. In 1986-1988, however, Lindberg's UMLS ideas were new to many in the informatics field and not very clearly expressed. The majority of potential users were not yet attempting to retrieve information from multiple disparate sources, let alone a mixture of evolving internal and external databases. There were few examples of knowledge artifacts intended primarily for use by system developers as opposed to end-users, and little experience with customizing multi-purpose resources for specific applications. Not surprisingly, the successful UMLS contractors came to the project with differing interpretations of its purpose and potential methods, and different ideas about the terminology problems and priorities NLM should address immediately.

Once decisions about the basic parameters for the initial UMLS Knowledge Sources were made in late 1988, Lindberg and others on the NLM UMLS Team began to publish clearer and more definitive statements about the UMLS goals and assumptions, contradicting some of the misconceptions then circulating.

> "The Unified Medical Language System (UMLS) project is … designed to facilitate the retrieval and integration of information from many machine-readable information sources, including descriptions of the biomedical literature, clinical records, factual databanks, and medical knowledge bases. The UMLS project is not an attempt to impose either a single standard vocabulary, a single standard record

format, or a single medical knowledge base on the biomedical community. The UMLS approach assumes that diversity will continue to exist and therefore seeks to provide products that can compensate for differences in the vocabularies or coding schemes used in different systems, as well as for differences in the terminology employed by system users." [14, p.475]

Three additional explicit UMLS assumptions reflected Lindberg's pragmatic views about system development in general. The first was a well-known Lindberg maxim, expressed in this instance as "information systems must be used if they are to improve," [14, p.475]. He expected the new and different UMLS components to begin as relatively simple structures and to go through iterative development with input and feedback from the intended users based on testing and use. "Complexity will be added in subsequent versions as actual use shows it to be necessary" [14, p.475]. The imperative for iterative development with user feedback dictated release of new editions of the UMLS components at least annually. Given their novelty, size, and initial lack of tooling, obtaining input on the early versions from system developers and researchers was difficult. In addition to free worldwide dissemination and internal testing and use, NLM employed various funding mechanisms to promote testing, use, and feedback. This was an unusual practice in the early 1990s, although it became more common later.

The second assumption, "effective information systems must interact with the end user," presupposes the presence of a user of any system employing the UMLS components to verify the interpretation of queries and resolve ambiguities beyond the system's understanding [14, p. 475]. Lindberg did not expect use of the UMLS components to enable information systems to produce perfectly relevant results, as if by magic, based on an initial user query. Early descriptions of how users might interact with systems that used UMLS knowledge imply a greater degree of iteration with individual users than actually became the norm after the arrival of the World WideWeb (WWW). Current systems employing UMLS or other resources to provide linked access to multiple information sources favor strategies designed to reduce individual user effort, although the user is still the final arbiter of what is relevant. These strategies include precomputed links among related information; established connections to specific information sources, e.g., via the Infobutton standard; and shaping current retrieval based on analysis of user search history.

The third assumption was "UMLS development will not be dependent on any projected or possible improvements in the basic information sources to which the UMLS will relate" [14, p. 476]. Lindberg applied this principle to other major NLM initiatives during his tenure. He viewed new and unanticipated developments as inevitable and was ready to take advantage of them when they occurred. He did not, however, commit major NLM resources premised on the future arrival of any specific development over which NLM had no control.

4. Setting General Parameters for the UMLS Metathesaurus and Semantic Network

Decisions about the scope and general structure of the UMLS Metathesaurus and Semantic Network emerged from an intense iterative process, informed by the work and opinions of all UMLS research groups [9, p.5-6]. Early statements about the UMLS project implied possible development of a new vocabulary to which existing terminology

systems would be mapped. While some UMLS-funded work explored structures for a new canonical representation of clinical concepts, Lindberg viewed development of a new clinical vocabulary as inappropriate for NLM ("The Library doesn't have patients"). Creating yet another biomedical terminology seemed counter-productive to the NLM team and too time-consuming as a first step toward the UMLS goal of facilitating retrieval of conceptually related information from multiple machine-readable sources.

Based on previous experience with processing words and terms from machine-readable texts and terminologies, the University of California, San Francisco (UCSF) UMLS Team proposed a different approach: "bootstrapping", or pre-computing, a draft Metathesaurus from existing terminologies and coding systems. The application of advanced computational methods to direct reuse of existing machine-readable vocabulary sources appealed to Lindberg. It struck the NLM Team as a more feasible, scalable, and still useful way forward, provided synonymy was confirmed or established among terms from the different vocabulary sources. In other words, the Metathesaurus would be organized by concept. Methods proposed by UCSF would help domain experts to achieve this. If one source asserted that two medical terms were synonyms or closely related, then those and other lexically similar terms could be collected into a single record for subsequent expert review.

Sample records illustrating the proposed methods and, importantly, a concept organization, were produced for review by all UMLS project participants. The sample records clarified the intent to include in the Metathesaurus all the terms and hierarchical categorizations for each concept from all its vocabulary sources, irrespective of conflicts within or among them. Each vocabulary's hierarchy, for example, was deemed essential to facilitate retrieval from databases indexed or encoded with it. Increasing the degree of unfamiliarity for those working on the project, at that time "concept-based" representations were not widely used in scalable information systems.

Many were skeptical about the value of a Metathesaurus with these parameters and adamant about the need for some level of consistent categorization of all concepts included. Based on their strong recommendations, NLM decided to create a separate UMLS Semantic Network, consisting of high-level Semantic Types or categories, e.g., Medical Device, Anatomic Abnormality, and the sensible relationships among them. Every Metathesaurus concept would be assigned at least one of the Semantic Types. This was an added task requiring domain expert review, but Semantic Type assignment proved to have major benefits for Metathesaurus construction and maintenance, as well as for use of the UMLS, e.g., in natural language processing (NLP).

With these decisions made, an NLM group led by Hole and Lexical Technology, Inc. (LTI), a firm formed by members of the UCSF UMLS team, focused on producing the Metathesaurus. McCray led the development of the Semantic Network, with input from all UMLS research groups. In this case, as in others, Lindberg did not expect perfection, but he did expect increased understanding of the problems involved, quickly produced first versions that showed some promise, and steady improvement in subsequent versions based on feedback from users. In the presence of all of these, he was willing to weather criticism from early users and ignore most comments from non-users.

5. Building the UMLS Metathesaurus

The production of the Metathesaurus was a "Big Data Science" project for its time, requiring substantial computing power for lexical matching and context representation and sophisticated large screen displays to assist domain experts in grasping the semantics and details of source vocabularies. The initial 1990 version had 64,123 concepts and 208,559 concept names from 7 vocabularies, thus dwarfing each of its components. Metathesaurus construction and maintenance was a bi-coastal operation with the NLM team in Bethesda, Maryland and the LTI Team in Alameda, California, so high-speed communications were also essential. At a time when it was unusual, LTI became an Internet node. This enabled sometimes overnight revision of Metathesaurus content when release deadlines loomed. In his dual roles as NLM Director and the first Director of the National Coordination Office of the High-Performance Computing and Communications (HPCC) (1992-1995), Lindberg funded, followed, and highlighted UMLS use of HPCC technology, which became more and more critical to Metathesaurus production as its size and complexity increased [15].

Typical for data science projects, "data wrangling" was a huge challenge for Metathesaurus creation and maintenance. At the time, LTI called it "source inversion" to denote the process of determining the internal semantics of each source vocabulary and transforming its "raw" machine-readable version into a common explicitly tagged representation for use in lexical matching and computing draft Metathesaurus entries. In current data science parlance, the development and ongoing maintenance of the Metathesaurus can be viewed as a largely successful effort to make terminology data more FAIR (findable, accessible, interoperable, and reusable) [16]. These themes were inherent in Lindberg's earliest statements about the UMLS.

All the "source vocabularies" for the Metathesaurus had content worth reusing, but the state of the art in machine-readable representation of terminologies was primitive. Technical formats ranged from simple word processing files to print tapes to databases. In some cases, a printed book was considered the authoritative version; some content visible in print to the human eye, e.g., conveyed by indentations or different type fonts, was difficult, if not impossible, to infer from the machine-readable version. Many sources lacked explicit metadata or explanatory documentation in any form. With the partial exception of MeSH, none had implemented formal change-tracking. As a result, a significant burden placed on Metathesaurus maintenance was the detection and interpretation of changes in new versions of the constituent sources and the invention of better change representation mechanisms.

Metathesaurus development and maintenance raised consciousness about the value of assigning permanent unique non-semantic identifiers, i.e., "the name that never changes", to concepts in terminologies and classifications. When Metathesaurus construction began, if vocabulary sources had unique identifiers, they generally were codes that conveyed the meaning of the concepts to which they were attached. Meanings of codes could change over time if the name changed. In extreme cases, a specific code might be retired and then later reused for a different concept. Codes might misrepresent the meaning of new concepts if inadequate "room" existed for creating new codes. Only one salient aspect of a concept was represented in a meaningful code, e.g., pulmonary tuberculosis as either a lung disease OR an infectious disease, but not both. Based on his experience, Lindberg was highly critical of the practice of relying on codes as the sole indication of biomedical meaning in electronic health data. He favored storage of

biomedical terms, as well as codes, to enable more accurate interpretation of current patients' data by health professionals and of longitudinal data for research.

As became evident, no biomedical terminology systems were strictly organized by concept prior to the production of the Metathesaurus. Under Schuyler's direction, NLM added concept organization, permanent context-free identifiers, and other features to MeSH in 1988/9 to, among other objectives, ease Metathesaurus production and maintenance. Most source vocabularies had one or more "entry terms" pointing to the preferred name or code associated with a concept, but did not express precise distinctions or relationships, e.g., synonymy, among them. Verifying and establishing synonymy among the names and codes in individual source vocabularies was therefore as essential to producing a Metathesaurus organized by concept as was establishing synonymy across different vocabulary sources.

NLM committed to ensuring that each source's view of the relationships among its terms was extractable from the Metathesaurus, i.e., "source transparency" [17]. By contrast, due to competing views of synonymy within its different sources, the Metathesaurus' own concept structure had to represent a single view. A pragmatic approach emerged. The most fine-grained authoritative distinction would "win" over larger-grained aggregates. In other words, if a distinction between two concepts mattered in some biomedical or health-related context, then there would be two concepts in the Metathesaurus [18].

End-user assessments of the coverage of early Metathesaurus versions prompted major revisions - thus proving Lindberg's rule, "use generates improvement." Metathesaurus file structure changed, multiple word and term indices were added, and from 1994 onward, UMLS releases included the SPECIALIST lexicon and lexical tools. Early experiments to determine whether the Metathesaurus embodied specific sets of terms produced variable and often irreproducible results. Often users' publications claimed that the Metathesaurus lacked certain specific content that was in fact present. Adding word, normalized word, and normalized string indices to the Metathesaurus files and including the lexical resources used to generate these indices in the UMLS release immediately improved the comparability of vocabulary matching results and provided the foundation for future tools that simplified UMLS use, e.g., MetaMap [19-20].

Lindberg always left a door open for changes in direction in the face of new knowledge and opportunities. Nevertheless, relatively early decisions about Metathesaurus scope, content, and semantics remain in effect today, despite enormous increases in its size [21]. The 2021 AA version contains 4.4 million concepts and 13,668,045 concept names from 218 vocabulary sources. Important enduring Metathesaurus characteristics include: a scope defined by the combined scope of its source vocabularies; organization by concept; permanent non-semantic concept unique identifiers (CUIs); assignment of high-level semantic types to all concepts; and inclusion and explicit attribution of each source's terms and relationships in a common fully specified format, irrespective of conflicts with other sources.

Precise attribution of the sources of content in the Metathesaurus gradually improved over successive versions [17]. This made change management more tractable. Many producers also made it a sine qua non for UMLS inclusion of their vocabularies (especially those with use restrictions). It supported accurate and efficient exclusion of vocabularies for particular applications and facilitated Metathesaurus updates.

6. Developing the UMLS Semantic Network

The UMLS Semantic Network consists of (1) a set of broad categories, or Semantic Types, e.g., "Pharmacologic Substance", "Disease or Syndrome", "Geographic Area", that provide a consistent categorization of all concepts represented in the Metathesaurus, and (2) a set of useful and important relationships or Semantic Relations that exist between the Semantic Types, e.g., "Causes", "Treats." The hierarchical or "Isa", relationship. e.g., "Geographic Area" Isa "Spatial Concept", enables Semantic Types to inherit properties from their ancestor Types. The most specific type applicable is assigned to each Metathesaurus concept. In an expression Lindberg liked, the Semantic Network is in essence a computer-readable representation of biomedical "common sense," to which each Metathesaurus concept is linked by virtue of its Semantic Type assignment [9].

The development of the Semantic Network differed from the development of the Metathesaurus in several respects. It was not a Big Data project: the first version had 131 Semantic Types and 34 Semantic Relations. Its structure was not novel: it was based on artificial intelligence (AI) theory and practice on knowledge representation for natural language processing (NLP). There was no direct reuse of existing content, but, in line with Lindberg's preferences, its new content was influenced by analyses of relevant "facts on the ground" by UMLS research teams. These included categories in the MeSH tree structures (MeSH has the broadest scope of the Metathesaurus source vocabularies) and relationships represented in clinical knowledge sources, NLP research, and MEDLINE queries and citation records. Importantly, the first public version of the Semantic Network reflected improvements made after a test involving preliminary Type assignments to 30,000 Metathesaurus concepts [22-23].

What was new about the Semantic Network and distinguished it from similar contemporaneous efforts was its very broad coverage [22]. Its scope had to support high level categorization of all concepts in the Metathesaurus source vocabularies. For example, MeSH encompasses a wide range of concepts, e.g., World Health Organization, Medicare, Buddhism, Civil Rights, Life Change Events, Cost-Benefit Analysis. As a result, the Semantic Network was the first "upper-level ontology" for the biomedical domain, with categories applicable to concepts in intersecting domains [24].

As with other UMLS resources, the plan was to add content and complexity to the Semantic Network only as use showed it to be necessary. Lindberg wondered whether the Semantic Network would eventually need more Semantic Types or more Semantic Relations [9]. As shown in the Semantic Network Archive, there was growth in the number of Relations and the number of relationships asserted between Semantic Types during the first decade of UMLS use, but changes have been relatively minor since that time [25]. The current version, stable since 2015, contains 127 Types and 54 Relations. Additions and deletions of Semantic Types cannot be made lightly given the downstream effect on Metathesaurus maintenance. There has been relatively little user demand for more granular Types. Instead, many users prefer to group Semantic Types, e.g., all types for health "problems", to aggregate concepts for various NLP and data mining tasks. In 2001, NLM added "Semantic Type Groups" to the UMLS release to meet this need [26].

Among many other uses, Semantic Types are a quick way to distinguish ambiguous terms, e.g. Sodium (Biologically Active Substance) vs. Sodium (Laboratory Procedure). The assignment of candidate Types to new additions to the Metathesaurus based on the purpose, e.g., disease classification, or hierarchy, e.g., neoplasms, of the source vocabulary avoids incorrect grouping of lexically similar, but semantically different

terms during Metathesaurus updates, thereby reducing work for expert reviewers. The number of under-specified concept names has diminished over time, e.g., "Cold" instead of "Cold Temperature," one of the many improvements in source vocabularies influenced by the UMLS project.

7. Incorporating the SPECIALIST Lexicon and Lexical Programs

In parallel with the early phases of the UMLS project (1986-1990), McCray and the NLM NLP group she formed developed SPECIALIST, a prototype system for parsing and accessing medical text. Lindberg had no specific guiding role on this effort beyond recruiting McCray to establish a linguistics research program at NLM and applauding the results. The SPECIALIST Lexicon and lexical tools were created to provide linguistic knowledge, i.e., lexical information, and rules of morphology, syntax, and semantics, "based on the assumption that systems combining domain knowledge with sophisticated linguistic analysis will lead to improved representation and retrieval of biomedical knowledge" [27, p.103]. Because biomedical language intersects with the standard language, the Lexicon encompassed general (standard) English lexical items, as well as biomedical domain specific lexical items [27]. In addition to other sources of general English and biomedical terms, the NLP group analyzed language in MEDLINE citations and abstracts to identify frequently occurring words and terms for inclusion in the Lexicon. They relied on MeSH as one source of domain knowledge, adding labels to the relationships in MeSH hierarchies which were subsequently incorporated into the Metathesaurus.

When the early versions of the UMLS Knowledge Sources were released, the NLM NLP group became active and sophisticated users, employing them to extend the capabilities and coverage of the SPECIALIST system and Lexicon and providing important feedback and assistance on useful UMLS improvements [28]. Experiments by external UMLS research teams also involved a range of automated lexical matching methods to map other vocabularies and free text to early versions of the Metathesaurus [e.g., 29-31]. The variable results of these experiments demonstrated the need to include word and term indexes in the Metathesaurus. Members of external UMLS teams, notably Columbia and LTI, encouraged NLM to release the SPECIALIST lexicon and lexical tools as part of the UMLS Knowledge Sources and to use them to produce normalized word and term indexes for the Metathesaurus. NLM added the Lexicon and lexical tools to the UMLS release in 1994 [19].

The SPECIALIST Lexicon and lexical programs were the first openly available and regularly updated biomedical lexical resources in English. Their release, both separately and as part of the UMLS Knowledge Sources, provided an unparalleled opportunity for research and development in biomedical NLP. Within a year of their addition to the UMLS Knowledge Sources compact discs, NLM made all the UMLS components available on the Internet from a UMLS Knowledge Source Server. The server had three different client interfaces: a Web interface for browsing and exploring, a command line interface for batch processing, and an application programming interface (API) to enable embedded calls to UMLS resources from external programs [32]. The new access methods made possible by the spread of HPCC technology, in combination with the addition of the lexical components, triggered substantial increases in use of the UMLS resources, particularly in NLP research and development.

8. Considering the Impact of the UMLS

In 1984, Donald A.B. Lindberg M.D. conceived the UMLS. Because it is both an evolving set of artifacts and a set of ideas, it is hard to find, over the ensuing nearly 40 years, a large biomedical information project that has not been influenced by it. Today, as a testament to Lindberg's foresight, the UMLS is infrastructure - heavily used, but not always cited [12,33-34]. As described here, it had no precedent, and, thus, initially, application developers, and their end-users, had difficulty applying it. But, as Lindberg often said, "Things that are used tend to get better." Slowly the field adopted either the UMLS artifacts themselves, its content, such as the synonyms, or its ideas, such as concept-based representations. While computers still struggle to "understand" biomedical meaning usefully, most would agree that Lindberg's vision and development approach enabled substantial progress in this important area.

The UMLS remains useful because Lindberg's 1984 expectations for the future in which it would operate proved to be highly accurate: exponential growth in biomedical and health data; great advances in computing and communications; increasing importance of molecular biology and genetics in research, knowledge discovery, and health care; greater patient interest in, and access to, health information; and no single standard language capable of meeting all biomedical and health needs, despite UMLS-aided progress toward clinical terminology standards [35]. The UMLS was initially ahead of its time and therefore ready for use when the future Lindberg envisioned arrived.

Endnote

[1] NLM released an experimental UMLS Information Sources Map (ISM) from 1991-1997. It ceased when Internet search engines and the World Wide Web changed many aspects of the problem of creating machine interpretable metadata for digital information sources. Greater integration and better discovery and linking mechanisms reduced the problem for NLM's own information resources. Nonetheless, it is an unsolved problem and continues to be the focus of current work, e.g., Alper BS, Flynn A, Bray BE, et al. Categorizing metadata to help mobilize computable biomedical knowledge. Learning Health Systems. 2021 DOI: 10.1002/lrh2.1027

References

[1] National Library of Medicine. Swearing-in ceremony: Donald A.B. Lindberg, M.D., Director, National Library of Medicine, October 11, 1984 [Internet] [cited 2021 July 8] Available from: http://resource.nlm.nih.gov/101629547.

[2] Departments of Labor, Health and Human Services, Education, and Related Agencies Appropriations for 1986: Hearings before the subcommittee on the Departments of Labor, Health and Human Services, Education, and related agencies of the House Committee on Appropriations, 99th Cong., 1st Sess. Part 4B, (857) (1985) (statement of Dr. Donald A. B. Lindberg, Director of the National Library of Medicine).

[3] Matheson NW, Cooper JA. Academic information in the academic health sciences center. Roles for the library in information management. J Med Educ. 1982 Oct;57(10 Pt 2):1-93..DOI:10.1097/00001888-198210000-00001---

[4] Blois MS. Information and medicine: the nature of medical descriptions. Berkeley: University of California Press; 1984. [cited 5 March 2021] Available from: http://resource.nlm.nih.gov/8400408

[5] Finkel AJ. Current medical information & terminology (CMIT): for the naming and description of diseases and conditions in practice and in areas related to medicine. Chicago, Ill.: American Medical Association; c1981.

[6] Lindberg DAB, Rowland LR, Buch Jr CR, Morse WF, Morse SS. CONSIDER: a computer program for medical instruction. 9th IBM Medical Symposium. 1968.

[7] Blois MS, Tuttle MS, Sherertz DD. Reconsider: a program for generating differential diagnoses. Proc Annu Symp Comput Appl Med Care. 1981 Nov 4 :263–268.
[8] Kingsland LC 3rd, Kulikowski CA. A scientific mind embraces medicine: Donald Lindberg's education and early career. In: Humphreys BL, Logan RA, Miller RA, Siegel ER, editors. Transforming biomedical informatics and health information access: Don Lindberg and the U.S. National Library of Medicine. Amsterdam: IOS Press; 2021.
[9] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Meth Inf Med 1993;32: 281-91.
[10] Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998 Jan/Feb;5(1):1-11.
[11] McCray AT, Miller RA. Making the conceptual connections: the Unified Medical Language System (UMLS) after a decade of research and development. J Am Med Inform Assoc. 1998 Jan-Feb;5(1):129-30. DOI: 10.1136/jamia.1998.0050129.
[12] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. DOI: 10.1093/nar/gkh061.PMID: 14681409
[13] Amos L, Anderson D, Brody S, Ripple A, Humphreys BL. UMLS users and uses: a current overview. J Am Med Inform Assoc. 2020;27(10):1606-11. DOI: 10.1093/jamia/ocaa08419
[14] Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. In: Kingsland, L.C. 3rd (ED.) Proceedings; Thirteenth Annual Symposium on Computer Applications in Medical Care, Washington, IEEE Computer Society Press, 1989; 475-480.
[15] Ackerman MJ, Howe SE, Masys DR. Don Lindberg, high performance computing and communications, and telemedicine. In: Humphreys BL, Logan RA, Miller RA, Siegel ER, editors. Transforming biomedical informatics and health information access: Don Lindberg and the U.S. National Library of Medicine. Amsterdam: IOS Press; 2021.
[16] FAIR principles. [cited 3 August 2020] Available from https://www.go-fair.org/fair-principles/ .
[17] Hole WT, Carlsen BA, Tuttle MS, Srinivasan S, Lipow SS, Olson NE, Sherertz DD, Humphreys BL. Achieving "source transparency" in the UMLS Metathesaurus. Stud Health Technol Inform. 2004;107(Pt 1):371-5.PMID: 15360837
[18] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993 Apr;81(2):217-22.
[19] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994 : 235–239. PMCID: PMC2247735
[20] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36. DOI: 10.1136/jamia.2009.002733.
[21] 2. Metathesaurus. UMLS Reference Manual. Bethesda, MD: U.S. National Library of Medicine, 2009.
[22] McCray AT. The UMLS semantic network. Proc Annu Symp Comput Appl Med Care. 1989 Nov 8: 503–507. PMCID: PMC2245676
[23] McCray AT, Hole WT. The scope and structure of the First Version of the UMLS Semantic Network. Proc Annu Symp Comput Appl Med Care. 1990 Nov 7: 126–130.
[24] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003 Feb; 4(1): 80–84. DOI: 10.1002/cfg.255 PMCID: PMC2447396
[25] UMLS Semantic Network Archive. [cited 12 May 2021]. Available from https://lhncbc.nlm.nih.gov/semanticnetwork/SemanticNetworkArchive.html
[26] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform. 2001; 84(0 1): 216–220.
[27] McCray AT, Sponsler JL, Brylawski B, Browne AC. The role of lexical knowledge in biomedical text understanding. Proc Annu Symp Comput Appl Med Care. 1987 Nov 4: 103–107. PMCID: PMC2245098
[28] McCray AT. Extending a natural language parser with UMLS knowledge. Proc Annu Symp Comput Appl Med Care. 1991: 194–198. PMCID: PMC2247522
[29] Huff SM, Warner HR. A comparison of Meta-1 and HELP terms: implications for clinical data. Proc Annu Symp Comput Appl Med Care. 1990 Nov 7: 166–169. PMCID: PMC2245515
[30] Cimino JJ. Representation of clinical laboratory terminology in the Unified Medical Language System. Proc Annu Symp Comput Appl Med Care. 1991: 199–203. PMCID: PMC2247523
[31] Miller RA, Gieszczykiewicz FM, Vries JK, Cooper GF. CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources. Proc Annu Symp Comput Appl Med Care. 1992: 86–90. PMCID: PMC2248100
[32] McCray AT. Razi AM, Bangalore AK, Browne AC, Stavri PZ. The UMLS Knowledge Source Server: a versatile Internet-based research tool. Proc AMIA Annu Fall Symp. 1996: 164–168. PMCID: PMC2233094

[33] Humphreys BL, Del Fiol G, Xu H. The UMLS knowledge sources at 30: indispensable to current research and applications in biomedical informatics. J Am Med Inform Assoc. 2020 Oct 1;27(10):1499-1501. DOI: 10.1093/jamia/ocaa208.
[34] Kim MC, Nam S, Wang F, Zhu Y. Mapping scientific landscapes in UMLS research: a scientometric review. J Am Med Inform Assoc. 2020 Oct 1;27(10):1612-1624. DOI: 10.1093/jamia/ocaa107.
[35] McDonald CJ, Humphreys BL. NLM and standards for electronic health records: one thing led to another. In: Humphreys BL, Logan RA, Miller RA, Siegel ER, editors. Transforming biomedical informatics and health information access: Don Lindberg and the U.S. National Library of Medicine. Amsterdam: IOS Press; 2021.