Appendix 2: Overview of the UMLS

The UMLS is populated from several other medical thesauri, such as Medical Subject Headings (MESH), ICD-9, SNOMED, and so on. Its contents center around concepts, terms, strings and words. A *concept*, such as hypertension, may be expressed in different ways: e.g., hypertension , high blood pressure disease and hypertensive vascular disease are different synonymous forms, or *terms,* that refer to the same concept. (Strictly speaking, a concept has a numeric ID only. However, a *preferred form* of the term can be used as the concept description: henceforth, when we refer to the concept in the context of a string match, we refer to its preferred form.) The same term may be expressed in multiple *string* forms through variations such as transposition of words, punctuation, or differences in case, person and tense, e.g., disease, hypertensive . Finally, each string is composed of one or more *words*. The UMLS contains numerous cross-reference tables that greatly ease the programmer s task. For example, the language-specific MRXW tables, which allow direct location of the IDs of all concepts containing a particular word, are key to the operation of many concept-matching algorithms.

A concept can belong to one or more *semantic categories* (e.g., pharmacologic substance, therapeutic procedure) and every term for a concept is tagged with the *ID of the source vocabulary* from which it was taken. This information allows researchers to create UMLS data subsets for special purposes.