3.1: Database Basics

What is a database

A database is an “organized collection of information.” The information in a database can be in any format, including texts, numbers, images, audios, videos, and many others (and combination of these), but this information must be “organized” for efficient retrieval. According to this definition, a database is not necessarily electronic (i.e., accessible by computers). For example, the collection of names in a phone book or address book may also be considered as a database, because the names are arranged (typically in alphabetical order) to make it easy to search for necessary information (e.g., phone numbers or addresses). However, in computer science and related areas, a database usually means an electronic database. Therefore, the term “database” in this module is used to mean an “electronic database”.

Primary vs. secondary databases

Databases are often categorized into primary and secondary databases.

- **Primary databases** contain experimentally-derived data that are directly submitted by researchers (also called “primary data”). In essence, these databases serve as archives that keep original data. Therefore, they are also known as archival databases.
- **Secondary databases** contain secondary data, which are derived from analyzing and interpreting primary data. These databases often provide value-added information related to the primary data, by using information from other databases and scientific literature. Essentially, secondary databases serve as reference libraries for the scientific community, providing highly curated reviews about primary data. For this reason, they are also known as curated databases, or knowledgebase.

It should be noted that the distinction between primary and secondary databases is not always clear and that many databases have the characteristics of both primary and secondary database. It is very common that a primary database curates its data with information drawn from secondary databases. In addition, because many secondary databases make their value-added information available in the public domain, data exchange and integration among databases very frequently occurs. As a results, virtually all data providers also becomes data consumers these days.

Data provenance

The term “data provenance” refers to a record trail that describes the origin or source of a piece of data and the process by which it entered in a database.[1] Simply put, data provenance deals with the questions “where the data came from” and “how and why the data is in its present place”. Although the data provenance information is critical in the reliability of a data source (and its data), this information is not easy to manage. In addition, information predicted in one database may not be appropriate for use in other databases, but may end up being integrated in them anyway. Therefore, databases need to document the provenance of the data and devise a way to notify users of that information. In turn, users should always pay attention to the data provenance issue when using a database.

References

1. Ram, S.; Liu, J. In *SWPM'09 Proceedings of the First International Conference on Semantic Web in Provenance Management* Washington, D.C. , 2009; Vol. 526, p 35.