# Information Science

Patrick Manning, *Methods for Human History: Studying Social, Cultural, and Biological Evolution* (New York: Palgrave Macmillan, 2020), 49–53.

## Prologue

The field of Information Science is quite new, having formed in the late twentieth century, though many of its constituent disciplines and methods came into existence much earlier. Expansion of information science has come especially because of the growth of computer technology and quantitative methods. This quick introduction to information science, a selection from Methods for Human History, gives examples of work in archives to hold data, ontology to define the groups of items within a domain, and datasets for organization and analysis of data. The full scope of information science broadens the understanding of "information" and applies its lessons to a wide range of disciplines.

## Essay

Information science congealed as a disciplinary focus in the last half of the twentieth century. It relied heavily on the expanding capacity of electronic computers but set computation in the context of analyses of information generally. Its fields include computer science, artificial intelligence, archival and library science, statistics, ontology, databases, simulation, and visualization. The field is still expanding and organizing itself. Applications and linkages to other disciplines are growing rapidly, including to health sciences, linguistics, management, museology, digital humanities, public policy, and social sciences. The discussion here touches on some but not all of its dimensions.

### Archives, Cataloguing Systems

Archives, databases, and cataloguing systems have developed impressively in many fields and have become remarkably accessible since the rise of the internet. The Human Areas Research Files (HRAF) brought an early advance in anthropology beginning 1947 and has been updated periodically.[1] The largest archive of medical and health research is the U.S. National Library of Medicine, which has grown under various titles since 1836.[2] In 2000, this institution opened PubMed, an open-access, online catalog of titles and abstracts for all medical and health literature, with more than 30 million citations and abstracts of peer-reviewed biomedical literature.

| |
|---|
| *Topic 4.6:* Archives<br>*Question:* **How can one locate topical resources within PubMed?**<br>*Theory:* Library science, health sciences<br>*Evidence:* Academic publications in health science and related fields<br>*Analysis category:* Define an ontology |

[1] Human Relations Area Files, https://hraf.yale.edu.

[2] National Library of Medicine (www.nlm.nih.gov): The Surgeon General's library, formed 1836 in Washington DC, grew steadily from 1865. It became the National Library of Medicine in 1962. Its electronic version began with Medline in the 1970s, which became the open-access PubMed in 2000. Another major resource is the Wellcome Library in London, at wellcomelibrary.org.

*Answer to the Question:* (Thought experiment) **Based on the online PubMed User Guide, develop a classification of topics of interest to you.**
*Presentation:* Identify relevant sections of the User Guide, note key terms and definitions, and prepare a one-page guide to finding publications of interest to you.
*Citation:* PubMed home, pubmed.ncbi.nlm.nih.gov/pubmed; PubMed User Guide, pubmed.ncbi.nlm.nih.gov/help

You are invited to explore PubMed through its User Guide and to prepare a brief list of search topics of interest to you. An example of a search term you can enter is "Tianyuan" or "Tianyuan Cave." This will lead you to articles on the remains of a man who lived 40,000 years ago and was found in a cave near Beijing. PubMed, once you have built up some practice, can be extremely useful. On the other hand, topical searches on Google are easy to submit and often bring up very useful results.

**Ontology**

Ontology is the categorization of groups and subgroups of items within a domain. Domain ontology is the categorization of items specific to an area of study. An upper ontology is a set of maximally general categories, such as continents within a geographical ontology.

Professional groups with a strong interest in sharing a common ontology for their area of work have developed domain ontologies including BORO Business Objects Reference Ontology; CIDOC Conceptual Reference Model, for cultural heritage; ISO 15926, International Standard for representation of plant life-cycle information; and UFO, Unified Foundational Ontology. As another domain ontology, a World-historical Gazetteer is under construction, listing global historical places and linking them to information on the places: see whgazetteer.org.

In this exercise, you are requested to categorize types of places in your home town. You may list individual places, but the main point is to identify the types of places and to suggest something about the relationship of the different places. You will encounter choices in categorizing places; then you will make the choices; finally you will look at the results of the choices you have made and see what you have learned. As an example of a very simple gazetteer, see the list of towns in Galicia (now Poland and Ukraine) remembered by Jewish emigrants, at www.geshergalicia.org.

*Topic 4.7:* Ontology
*Question:* **What should be the structure of a geographical ontology for your home town?**
*Theory:* Ontology
*Evidence:* Your own information on places in your home town.
*Analysis category*: Define an ontology.
*Answer to the Question:* (Thought experiment) **Identify places and types of places in your home town, and prepare an ontology.** It should organize places by type and scale, showing how the types are linked to one another.
*Presentation:* A one-page list of categories (with an example for each category) showing their links.
*Citation:* World-historical gazetteer, www.whgazetteer.org

**Databases**

Databases or datasets, generally stored in electronic form, are the results of collection, organization, displaying, analysis, interpretation and presentation of data. They are usually organized as two-dimensional spreadsheets, with variables listed in one axis, cases observed on the other axis, and with the observations of data in the cells linking variables and cases. Data may be collected from historical research in past documents, from current data such as online records, or from data mining and sampling of data from large data archives. Data must be verified for accuracy and consistency before analysis. If there are missing observations on variables, one faces choices on whether to neglect the variable or to simulate the missing data. Datasets may be constructed to fit with variables

in a theory in order to test theoretical relationships. Analysis of datasets can explore the nature and significance of relationships among variables.[3]

The Seshat database has been constructed for comparative and global analysis of societies since the beginnings of agriculture, expanding earlier such datasets.[4] In this analysis of numerous societies over time, the researchers ask whether the societies largely resemble each other, or whether they tend to fall into two or more categories. The method of analysis is *principal component analysis*, a statistical technique comparing multiple observations and using an algorithm to simplify them into one or more categories. In this case, most of the variation in the database can be explained by a single component, i.e., as variations on a single model. As the authors argue, "We found that different characteristics of social complexity are highly predictable across different world regions."

*Topic 4.8:* Databases
*Question:* **Do human societies from around the world exhibit similarities in the way that they are structured? Do they show commonalities in the ways that they have evolved?**
*Theory:* Authors sought to select data that were as "theory neutral" as possible.
*Evidence:* For polities selected from 30 "natural geographic areas" of the world, data were collected on 51 variables, aggregated into nine "complexity characteristics."
*Analysis category:* Verify a hypothesis.
*Answer to the Question:* **Yes. "Our analyses revealed** that these different characteristics show strong relationships with each other and **that a single principal component captures around three quarters of the observed variation**."
*Presentation*: The article describes the process of sampling (among regions and variables), an ontology for aspects of human society, and statistical procedures (focusing on principal component analysis).
*Citation:* Seshat, http://seshatdatabank.info; Turchin, et al., "Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization," PNAS 115 (2): E144-E151.

[3] For instance, the Caribbean inequality database, discussed earlier in Topic 1.5, has a different structure and a different purpose.

[4] Seshat: Global History Databank (seshatdatabank.info).