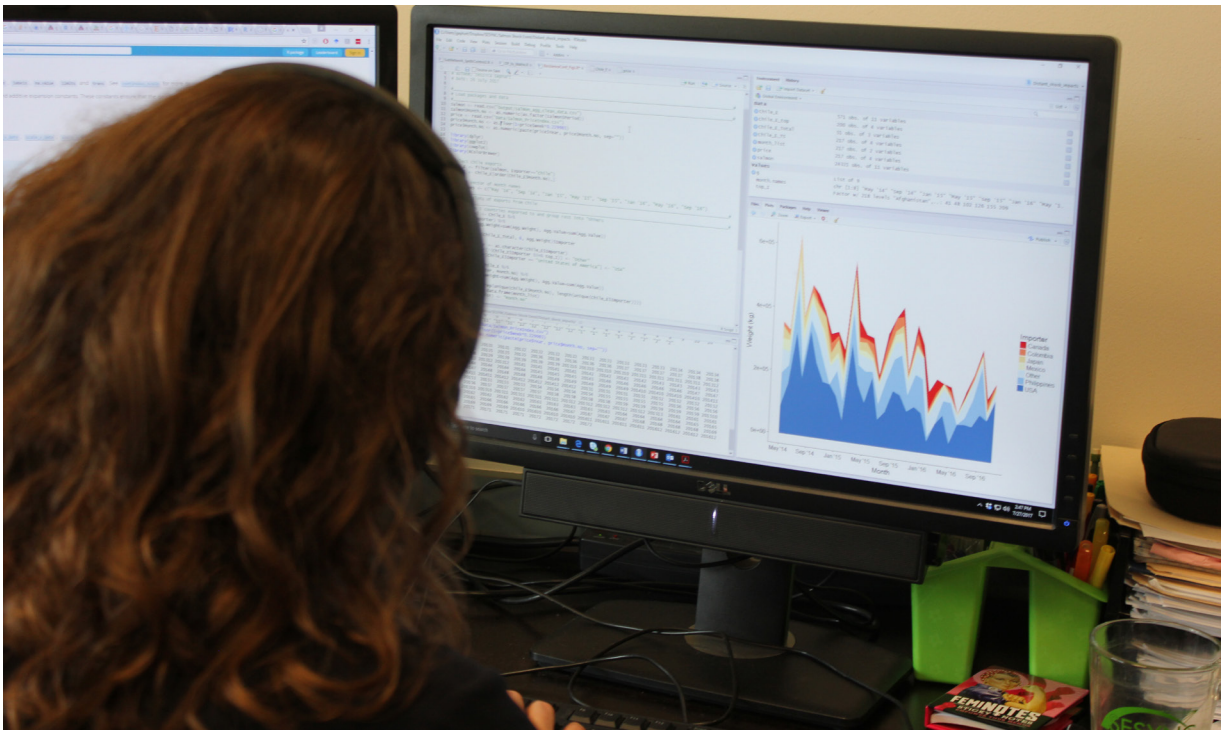
NATIONAL SOCIO-ENVIRONMENTAL SYNTHESIS CENTER



Explainer: Quantitative Synthesis Methods for Sustainability: Data Integration

By Margaret Palmer, SESYNC, and Kelly Hondula, SESYNC | August 26, 2022

Climate change, world hunger, and water insecurity are only a few of the major sustainability challenges the world faces today. Researchers seeking to address these issues require biophysical, social, economic, and technological data. However, these problems are too urgent to rely solely on the collection of new data—from experiments, sampling campaigns, interviews, policy evaluations, or technological traits—for answers. As a result, researchers are taking advantage of the plethora of data already in existence and using quantitative methods to integrate multiple data sets to address a broad range of sustainability questions. This work can involve integrating so-called "big data," such as terabytes of environmental data or information on how humans interact with or react to their surroundings; however, it can also involve integrating information from many small but highly heterogeneous sources of information such as an individual's results in a small spreadsheet.

Informatics and a Few Key Terms

The integration of data to address a problem involves many steps and processes including data discovery, collection, storage, computational processing, visualization, and interpretation of results. Informatics is the general term used to describe data-centric approaches to solving problems. It is a vast topic involving concepts and practices from multiple disciplines including computer science, statistics, data science, computational social science, and many other domains (e.g., ecology, genetics, health and medicine, cognitive science, and geography). Books have been written on each of these topics and there are even

entire academic programs devoted to them. For this reason, we only briefly define a few key terms that are essential. See also SESYNC's Database Principles and Use lesson on GitHub for more information.

Metadata – Metadata is information about the structure and content of a dataset that is essential for data discovery and re-use. It can adhere to international standards, some of which are domain specific. Metadata often descibes who collected the data, where, and how; it should be stored in a project's data folder or repository.

Relational Database – A relational database is a collection of data items in one or more tables that can be related to one another using a common attribute. Each column of a table contains information of the same type (e.g., a number or date), and each row contains information about the same entity (e.g., a person, household, or geographical unit).

Data Harmonization – Data harmonization is the process of building a composite dataset after ensuring data are in a consistent, standardized format. Often this process involves converting data to common units, but sometimes, data on the same topic have been collected using different methods or at different scales, and thus researchers must use assumptions, statistical modeling or other tools to make them comparable.

Programming Languages – Programming languages—including R, Python, SQL, and Stata—are commonly used to access and manipulate databases in an efficient and reproducible manner.

Structured vs. Unstructured Data – Structured data is highly organized, formatted to a set structure before being stored, and most often quantitative in nature. It's organized in columns and rows, easily processed, and machine-readable (e.g., a spreadsheet of temperature, crop yields, and prices). Unstructured data includes diverse types of data with no predefined structure, and it is often qualitative (e.g., videos, tweets, images, text, and files like PDFs).

Version Control Git and Github – Git is a distributed version-control system that's free and open-source and designed to handle everything from small to large projects with speed and efficiency, while Github is the platform for doing version control and for collaborating on projects. For more information see SESYNC's guide on Resources to Help You Learn GitHub Pages.

Resources

SESYNC provides links to data science resources and data science lessons but so do many other well-established organizations. Because SESYNC's Git-based resources will no longer be updated after March 2022, we recommend you check out resources from sites like Earth Lab and DataOne. Additionally, Software Carpentry has well-developed lessons and Data Carpentry hosts workshops and online lessons on introductory computational skills needed to manage and analyze research. Additionally, Broman and Woo (2018) provide an especially useful article for beginners: "Data Organization in Spreadsheets."

For researchers, you should also be aware of the need to make your data and metadata publicly available. There are number of organizations that facilitate data curation and re-use of data; here are a few: the Environmental Data Initiative (EDI); Dataverse (open-source data repository software) and associated customized Dataverse institutional collections like Harvard Dataverse; the Inter-university Consortium for Political and Social Research (ICPSR). Many of these organizations, such as ICPSR, provide training in data access, curation, and analysis methods for the research community.

This work was supported by the National Socio-Environmental Synthesis Center (SESYNC)
under funding received from the National Science Foundation DBI-1639145.