There is an active coalition of social science data archives internationally. Founded in 1967, the UK Data Archive (UKDA) is curator of the largest collection of digital data in the social sciences and humanities in the UK. The Inter-university Consortium for Political and Social Research (ICPSR), established in 1962, is an integral part of the infrastructure of social science research in the US.

The Social Science of Data Sharing: Distilling Past Efforts

Metadata matters, or retaining the context and relationships

The enhanced metadata standard, Data Documentation Initiative (DDI), for social science data, documents data elements and their relationships in micro-, meso-, and macro-levels (both explicit and implicit) to provide the information that others require to simply locate a dataset or enable replication studies. There is a range of practices in the social sciences from depositing data within institutional repositories, and data professionals wish to cite their data source especially if an online data archive as a condition of grant award. Government data collections often fostered data sharing, or what we would now recognize as digital data collections created by government or individual researchers offered rich resources for the original intent of the investigation as well as for the original investigators. Data collections have developed from different domains, such as demographic and health observational data to look at the human dimensions of climate change. This history of practice over the last 40 years, has not been without problems, as the cost of data sharing requires considerable investment and overreaches in creating archival and curatorial infrastructure.

Metadata is implemented at the start of a research project. A number of projects are underway to automate metadata generation and capture at least in part and provide incentives for data archivists, data librarians, and social scientists, drawn from a variety of settings, including data archives, statistical agencies, research centers, libraries, academic departments, and non-profit organizations. The diagram illustrates how ASSIST helps bridge the interests and concerns of three communities.

Appraisal activities, or deciding what needs to be saved for future use

Collection development policies and procedures that include appraisal guidelines have been routinely used in traditional archive and library environments. Similarly, social science data archives have developed well-established criteria for appraisal in order to prioritize the effort required to support data curation. Long-term preservation of social science data became more prevalent as the need for digital data preservation became more urgent. However, problems associated with differing metadata representations and challenges associated with identifying unique records in datasets are still present. Data in social science research is often complex and therefore requires more detailed description. The Inter-university Consortium for Digital Social Science Data (ICPSR), founded in 1967, seeks to improve the description of digital social science data through a process of appraising and selecting digital social science data.

The Inter-university Consortium for Digital Social Science Data (ICPSR) is a not-for-profit research organization that facilitates access to its digital social science data. The ICPSR provides access to social science data through its own online data repository, which is called the ICPSR Data Archive (ICDA). The ICPSR Data Archive contains over 14,000 datasets and more than 250,000 variable definitions. The ICPSR Data Archive is organized by discipline and subject area, and includes datasets from over 170 countries. The ICPSR Data Archive also includes data from government agencies, private organizations, and individual researchers.

Social Data Sharing: Distilling Past Efforts

The first social data sharing program was established in the United States in 1967. The ICPSR was founded in 1967 and is the oldest existing social science data sharing program in the United States. The ICPSR is a non-profit organization that provides access to social science data through its own online data repository, which is called the ICPSR Data Archive (ICDA). The ICPSR Data Archive contains over 14,000 datasets and more than 250,000 variable definitions. The ICPSR Data Archive is organized by discipline and subject area, and includes datasets from over 170 countries. The ICPSR Data Archive also includes data from government agencies, private organizations, and individual researchers.

With the rise of mainframe-based statistical software packages such as SPSS in the late 1970s, analysis of social science data became more prevalent and standardized. Input and output formats made long-term preservation of datasets easier. But it was not until the rise of personal computers in the 1980s that data curation became more sophisticated. With the development of data curation software, such as the Data Documentation Initiative (DDI), the need for standardized data curation became more apparent. The DDI is an international standard for describing digital social science data. The DDI provides a framework for describing social science data, and it is used by social science data archives to ensure that data are described in a consistent and interoperable way.

Creating metadata records of this magnitude is not trivial. A range of practices in the social sciences from depositing data within institutional repositories, and data professionals wish to cite their data source especially if an online data archive as a condition of grant award. Government data collections often fostered data sharing, or what we would now recognize as digital data collections created by government or individual researchers offered rich resources for the original intent of the investigation as well as for the original investigators. Data collections have developed from different domains, such as demographic and health observational data to look at the human dimensions of climate change. This history of practice over the last 40 years, has not been without problems, as the cost of data sharing requires considerable investment and overreaches in creating archival and curatorial infrastructure.


Incenting citation practices

Researchers should cite computer files for the same reasons they cite traditional resources. A proper citation gives credit to the original author and also provides the information that others require to simply locate a dataset or enable replication studies. There are a number of steps that data archivists and government agencies providing bibliographic citations for the data they disseminate can take to ensure that the data are cited. For example, data archivists can provide links to the original author’s website or provide a persistent identifier for the dataset. Data archivists can also provide a data citation template that includes the data source’s name, the data version number, and the date the data were accessed. These steps can help ensure that the data are cited in a consistent and interoperable way.

Online cross-tabular analysis of a UKDA survey using the Naster tool.

Variable-level XML markup using the DDI DTD (Data Resource Centre, University of Guelph)

The enhanced metadata standard, Data Documentation Initiative (DDI), for social science data, documents data elements and their relationships in micro-, meso-, and macro-levels (both explicit and implicit) to provide the information that others require to simply locate a dataset or enable replication studies. There is a range of practices in the social sciences from depositing data within institutional repositories, and data professionals wish to cite their data source especially if an online data archive as a condition of grant award. Government data collections often fostered data sharing, or what we would now recognize as digital data collections created by government or individual researchers offered rich resources for the original intent of the investigation as well as for the original investigators. Data collections have developed from different domains, such as demographic and health observational data to look at the human dimensions of climate change. This history of practice over the last 40 years, has not been without problems, as the cost of data sharing requires considerable investment and overreaches in creating archival and curatorial infrastructure.