Kat Fritz
March 6, 2023
LIS 545 B

Repository Profile

Repository: Harvard Dataverse

Introduction

Harvard Dataverse is an open-source research data repository platform where researchers can share, preserve, explore, and analyze research data. The platform is operated by the Institute of Quantitative Social Science (IQSS) at Harvard University. The repository offers free consultations and assessments as well as tiered data curation services researchers can purchase. The data are organized into dataverses containing datasets which hold descriptive metadata, data, and code. Dataverses can also be nested into other dataverses.

I chose Harvard Database because it is a general-purpose data repository. The dataset I chose, "Caffeine Content of Drinks", focuses only on caffeinated beverages, and I was unable to find an appropriate topical repository. Though I had looked at the repository FoodB as an option, it only contains unbranded food items. The dataset does contain some entities like simple *green tea*, but most entities are branded items, like *Red Bull*. Additionally, FoodB is closed for submissions while Harvard Dataverse is open to submissions.

Collection Scope and Eligibility for Data Contributions

Anyone is eligible to contribute data from all domains if they have the appropriate permissions. Though there are no stated limits to what can be deposited, the repository encourages the submission of data that is relevant to research and scholarship. Harvard Dataverse does not have any specific limits related file formats, accepting data in various formats, including spreadsheets, images, and text files. The only limits are data sizes; file uploads are permitted up to 2.5GB and

dataset uploads less than 1TB. However, the repository will make exceptions for Harvard researchers.

Data and Metadata

Researchers and other users can access the User Guide if they need any guidance on using the repository. For submitting data, users can visit the Dataset + File Management page of the guide for help on what should be in the Submission Information Package (SIP). The repository recommends that submitters provide as much information about the dataset as possible, to facilitate discovery and reuse by others. For instance, the SIP should include all data files, the title, a readme file, licensing information, and other metadata. There are three levels of metadata: citation metadata (general metadata), domain specific metadata, and file-level metadata. Domain specific metadata is "currently for Social Science, Life Science, Geospatial, and Astronomy datasets" (*Data + File Management*). For a dataset to receive a DOI and Data Citation, the required metadata are the title, authors/contributors, description, contact email, and subject. Requiring these fields ensures datasets align with FAIR principles: findable, accessible, interoperable, and reusable.

The Harvard Dataverse does not require metadata to be submitted in any specific structure or adhere to a specific standard. Instead, it supports many metadata schemas, including Citation Dublin Core, various DDI, Datacite 4, JSON, OAI_ORE, OpenAIRE, Schema.org, and JSON-LD. Currently, there is also an experimental workflow allowing data depositors "to create and deposit Differentially Private (DP) Metadata files, which can then be used for exploratory data analysis. This workflow allows researchers to view the DP metadata for a tabular file, determine whether or not the file contains useful information, and then make an informed decision about whether or not to request access to the original file" (*Data + File Management*).

Kat Fritz
March 6, 2023
LIS 545 B

Harvard Dataverse displays metadata using the Data Documentation Initiative (DDI) standard (*Preservation Policy*). DDI provides a structured way to present metadata, which makes it easier for users to understand and compare data across different studies. This ensures that the metadata is consistent, well-structured, and machine-readable. However, because the metadata fields are so customizable, not all datasets will adhere to DDI standards even for the interface's display.

Data Access Mechanisms

Researchers pick if they want to allow their data to be open to the general public or restricted. They also choose specific terms of access or allow users to request access if the data are restricted. Even if data are restricted, Harvard Dataverse makes the metadata searchable so users can find the research. Any data contained in the repository are extremely accessible as a login is not required to download data. However, if you want to deposit data, you need to create an account.

Harvard Dataverse provides multiple access mechanisms for data. Users can download data directly from the repository, access data via APIs, or use third-party tools such as R or Stata to access data. Additionally, the repository provides database search and filtering tools to help users find the data they need. This means that users can choose the access mechanism that best suits their needs and preferences by selecting specific *File Tags* in the advanced search.

Because the repository focuses on accessible data, when a tabular dataset is submitted, Harvard Dataverse automatically creates a TAB file that is offered alongside the original file format. Users can download all files in a dataset by clicking on "Access Dataset" to receive the Dissemination Information Package (DIP). Then, a user can download a ZIP file with the author's

original folder layouts or a TAB file for the archival format. "Access Dataset" also allows users to use integrations like Binder to analyze and compute data.

Conclusion

Harvard Dataverse is an open and accessible repository that provides a platform for archiving, sharing, and finding research data. Its submission requirements are clear and straightforward, and it provides several access mechanisms for users to download and reuse data. Its openness to a variety of file formats and metadata standards ensures submitted data are discoverable and accessible to a wide range of users.

Kat Fritz
March 6, 2023
LIS 545 B

References

Data + File Management, 2023. Retrieved February 28, 2023, from https://support.dataverse.harvard.edu/harvard-dataverse-preservation-policy

Harvard Dataverse, 2023. DATACC. Retrieved February 26, 2023, from https://www.datacc.org/en/warehouses/harvard-dataverse/

Preservation Policy, 2023. Retrieved March 2, 2023, from https://support.dataverse.harvard.edu/harvard-dataverse-preservation-policy