Guidelines for Dataset Collection Development

Last updated: April 3, 2024

The Carnegie Mellon University Libraries collects datasets and data products to support research and teaching at CMU. The following provides guidelines to assist in the decision-making process when considering datasets and data sources for the collections.

In the context of these guidelines, datasets and data sources refer to data, broadly defined, produced for purchase by non-CMU entities including commercial providers, government agencies, trade associations, and other organizations. Datasets require additional software to view and analyze, such as Excel or a statistical program like SAS or SPSS, and are typically purchased in the form of files on a DVD or accessible via FTP. Data sources refer to internet-based data that is available for purchase or subscription and includes access to an online platform that facilitates data access and viewing (with the possibility of some downloading/exporting). Data sources can also refer to bulk textual data for use in text mining applications or large image/video datasets prepared for use in computer vision and image processing research.

The following should be taken into consideration in decisions about purchasing or subscribing to a new dataset or data source:

- **Scope and Relevance**
    - Datasets with a broad, cross-disciplinary subject appeal to the CMU research community, supporting the research and learning mission of the institution, should be prioritized.
    - The value of the historical data, the uniqueness of the data, its availability elsewhere, and geographic scope should all be considered.
    - Note whether the value of a dataset is likely to increase or decrease over time.
    - In the case of longitudinal data, consider whether there is an additional cost to updating/refreshing the dataset over time, whether such updates are automatic or require additional purchase requests, and whether a lack of an update will impact the value of the dataset.
- **Storage and Accessibility**
    - Datasets should comply with the Libraries' existing storage capabilities.
    - Datasets provided via DVD or FTP should be made available via password protected, online access if possible. Thus, the size of the data should be taken into consideration and the Library Information Technology team should be consulted to determine the availability of server space as needed. In addition, licensing agreements should ensure that the data can be made available in this way (see Terms of Use below).

    - For data products, ongoing access to backfiles should be determined in cases of expired license agreements and/or removal of the product from the market.
    - For data products, remote access via IP authentication or Andrew ID is preferred over password or individual account access. Access for multiple simultaneous users is preferred over limited-use access.
    - Note that the Libraries cannot support the provision of confidential or restricted data such as data subject to HIPAA protection, or consumer credit data.
- **Cost**
    - As with other collection development decisions, cost should be taken into consideration and based on current budgetary constraints and the perceived value of the dataset or data source to the CMU community.
    - Purchases jointly made with a researcher or department can be considered in cases of Libraries budgetary constraints.
    - All new purchase requests should go through the Head of the Technical Services for approval.
- **Quality**
    - The supplier of the data and the data itself should be vetted for accuracy, descriptive metadata, provenance, authoritativeness, and completeness.
    - Frequency of content updating should be taken into consideration when relevant.
    - Vendor reliability should be assessed including responsiveness, accessibility, and the availability of technical support.
    - In cases where data is made available via an online platform that facilitates data access and viewing, the quality, accessibility, and reliability of the platform itself and its user interfaces should be assessed for usability and ability to meet researcher needs. The platform should tend towards “seamless access” for our users to the greatest extent possible.
- **Format**
    - Data should be provided in a (preferably open) format that can be supported by the library and used by the researcher.
    - Data should be readable and manipulable in commonly used statistical software such as SAS, SPSS, and R.
- **Terms of Use**
    - Datasets purchased should be institutionally accessible to all faculty, students and staff, even if they are offsite, and the terms should allow for sharing of the data in a password protected, online environment. If appropriate, it should be made clear to researchers that the data should not be shared with non-CMU affiliates.
    - Terms should be in accordance with those for other electronic resource purchases made by the Libraries.
    - Consider fair use and the rights of scholars to data derivatives.

    - Datasets that require non-disclosure agreements or other limiting agreements should be considered with caution, and special accommodations may need to be made to store the data on a restricted access server.
    - In cases that require a data use agreement, the Office of the General Counsel or Office of Sponsored Programs should be consulted.
- **Documentation**
    - Datasets and data sources should include adequate documentation, such as manuals, data dictionaries, or codebooks, including relevant metadata.
    - Consider the language of the data documentation.

APIs (Application Programming Interface)

Related to the collection of datasets is access to APIs. APIs provide a mechanism to request and translate data from a source, in a way that allows automation and updating. Data providers sometimes make APIs freely and openly available or free upon request to existing subscribers. Some APIs are available for an additional fee or are customizable to a researcher's specific needs.

The Libraries can provide access to APIs via existing subscriptions or can serve as a liaison between a data provider and a researcher to facilitate customized APIs. Selectors should take the following into consideration when assisting researchers with accessing APIs:

- Is the API necessary to access the data needed by the researcher?
- Is the API freely available through existing subscriptions? Work with our vendor representatives to determine if there are available APIs to meet the researcher's needs.
- If the API is only available for an additional fee, does it need to be customized or is there a pre-existing API that will meet the researcher's needs?
- Does the vendor provide support for working with the API?
- If there is a cost associated with the API? The selector should weigh the costs with the potential benefits for users beyond the requesting researcher, and should verify the availability of the API beyond a single researcher. APIs that are customized or not available beyond a single user should be paid for by the researcher.