# Background

## Data bundling

*Public data* is data shared without filing a data request. Interested data-reusing researcher can obtain the data within minutes. Data dictionary can be referenced with URL that works without logins.

*Restricted data* is data that requires a data reusing researcher to file request. Also may be referred to as “request-walled data”. URL may not be possible to obtain or may only work for logged-in users with approved request that includes the study in question.

Example

* Public data contains only human friendly data dictionary
* Advanced and computer-friendly data dictionaries may be bundled in the restricted data

## Stages of data re-use

* **Review stage**
* Manual or computerized – researcher uses manual human review to assess whether to further work with the data
* Mode
* Based on data dictionary (no access to data)
* Based on preview of the data (either demo dataset, or partial or full dataset)
* **Data request stage** – researcher files a formal request (with varying level of detail required by each platform) and waits for approval.
* **Data use stage** – data is loaded into analytical platform. May occur via API.
* Single study mode - procedure is specific to an individual study
* In batch mode - procedure that accesses multiple studies at the same time

## Repository infrastructure

* *Simple repository* is a repository that does not internally load all data into a single data platform. Typically, each study can separately and independently provide.
* Simple repository may still enforce strong metadata policies that still require to keep each study independent, but enforce some common requirements. (e.g., specific format for metadata)
* *Common metadata platform repository* is a platform that enforces stricter harmonization across included studies. At a central coordinating center, each study is loaded into a single data platform. Individual study exports are standardized by using an export feature within the central platform.
* Example: Project Data Sphere SAS-based cloud platform that includes all platform studies pre-loaded within an analytical platform.
* Examle: Vivli analytical platform

## Recommendations (data dictionary)

* Sharing a data dictionary in a format that is human friendly and readable facilitates human review; however, in addition to *human-friendly format*, a data reusing researcher who proceeds to actual analytical use of the data in a statistical platform, will have to tediously look up many data elements abbreviations (or other identifiers) repeatedly. A data re-using researchers who loads the data into a statistical platforms would benefit from a *computer-friendly format*.
* Consider that data re-using researcher may be using a different statistical platform. Avoid using SAS-specific format for sharing data dictionaries. Consider formats that can be loaded to at least two or ideally multiple platforms.
* Sharing data via REDCap. REDCap has the ability to convert data into multiple output format. This functionality is, however, enabled by providing the data, in the first place, in a neutral format (when loaded into REDCap).
* Relevant Data Dictionary standards are:
* REDCap CSV (and zipped) based format
* CDISC Define.XML standard
* R: <https://github.com/DataDrivenInc/R4DSXML>
* SAS: <http://support.sas.com/documentation/cdl/en/clinstdtktug/66870/HTML/default/viewer.htm#titlepage.htm>
* Dublin Core, ISO/IEC 11179 and others listed at <https://en.wikipedia.org/wiki/Metadata_standard#Available_metadata_standards>
* Unofficial: Share data dictionaries in a format that you would wish to receive when you need external data. (do to others as you would have them do to you)

## Subset of elements

PERSON (id, sex, year of birth (or age at index event)

VISIT\_OCCURRENCE (id, when)

MEASUREMENT (weight)

CONDITION\_OCCURRENCE (hiv infected dx)

## NOTES

Size: 50+ patients

Recent: Ongoing as of 2018-06-30 or complete date between July 2013-June 2018. (last 5 years)

R define.xml package <https://github.com/DataDrivenInc/R4DSXML>