CUNY J+/PMN skills workshop series

Handout: Most common data formats and concepts

URL of document: http://bit.ly/commonsdataformats

The world of data and computing is full of data formats, acronyms and concepts used in specific areas. Here is a useful list of terms you can come back to whenever you have doubts.

List of terms (A to Z)

Application Programming Interface or API

A way computer programs talk to one another. Can be understood in terms of how a programmer sends instructions between programs. For data, this is usually a way provided by the data publisher for programs or apps to read data directly over the web. The app sends the API a query (question) asking for the specific data it needs, e.g. the time of the next bus leaving a particular stop. This allows the app to use the data without downloading the whole dataset, saving bandwidth and ensuring that the data used is the most up-to-date available.

Bulk

Data is available in bulk if the entire dataset can be downloaded easily and efficiently to your computer.

CSV

CSV stands for 'Comma-separated values'. It is a standard format for spreadsheet data and is widely used in the public and private sectors to produce datasets. Data is represented in a plain text file, with each data row on a new line and commas separating the values on each row. As a very simple open format it is easy to consume and is widely used for publishing data. You can import CSV files into Excel.

Data cleaning

Processing a dataset to make it easier to consume. This may involve fixing inconsistencies and errors, removing things like formatting, using standard labels for row and column headings, ensuring that numbers, dates, and other quantities are represented appropriately, etc.

Data collection

Datasets are created by collecting data in different ways: from manual or automatic measurements (e.g. weather data), surveys (census data), records of decisions (budget data) or ongoing transactions (spending data), aggregation of many records (crime data), etc.

Database

Any organised collection of data may be considered a database. In this sense the word is synonymous with dataset. Also used to talk about a software system for processing and managing data, including features to extend or update, transform and query the data. For example: SQL or PostgreSQL or Microsoft Access.

Geodata

Any dataset where data points include a location, e.g. as latitude and longitude or another standard encoding. Maps, transport routes, environmental data and many other kinds of data can be published as geodata.

Identifier

The name of an object or concept in a database. An identifier may be the object's actual name (e.g. 'Philadelphia' or a Philadelphia zip code), or a word describing the concept ('population'), or an arbitrary identifier such as 'XY-1' or HOME_ORG, that makes sense only in the context of the particular dataset.

Machine readable

Data in a data format that can be automatically read and processed by a computer, such as CSV.

Metadata

Is the information about a dataset such as its title and description, method of collection, author or publisher, area and time period covered, license, date and frequency of release, record layout or codebook of a database or dataset. Always look for the metadata and request it when you do a FOIA request or any type of request involving datasets.

PDF

Portable Document Format is a file format intended to produce electronic paper. Government and company officials love it because data in PDF files is not machine-readable. It's a format used to lock the data in it. They use PDF, mostly, because they believe it safeguards data provenance and prevents others from forging the information in the documents. But as journalists we want the data trapped in PDF using tools like Tabula and Cometdocs (your newsroom has access to this software. You can ask Dylan Purcell).

Query

A type of question accepted by a database about the data it holds. A complex query may ask the database to select records according to some criteria, aggregate certain quantities across those records, etc. Many databases accept queries in the specialised language SQL or dialects of it. A web API allows an app to send queries to a database over the web.

Record Layout (or codebook or data dictionary)

Datasets and databases include id's and codes (numbers or acronyms or a mix of numbers and letters) adopted by their creators to reference the categories of the data. The most common code you will see is in the header. It can be something simple like TITLE_DESC which stands for Title Description; or ORG_LEVEL for Organization Level. This two are from

the School District of Philadelphia (SDP) data. A clean organized version of it can be found at the Philly.com Data Hub's "Philadelphia School District Payroll Search". The original source is the School District of Philadelphia website that maintains the database and publishes it in partnership with Open Data Philly. The data is available in the zip file SDP Employee Information (.ZIP). If you download it and open the **SDP Employee Information (Jan 2017)** you will be able to read the metadata and record layout of the data, like:

- This set of data was extracted on December 31st 2016, and includes every active employee of the School District of Philadelphia as of the extract date.
- Employee_information.csv provides basic information on every employee of the School District of Philadelphia.
- In this case the record layout is a simple text:
    - LAST_NAME - Employee's last name.
    - FIRST_NAME - Employee's first name.
    - PAY_RATE_TYPE - The pay rate type for the employee, either SALARIED, DAILY or HOURLY.
    - PAY_RATE - The pay rate for the employee. If the pay rate type is salaried, then the annual salary (minus any benefits). If the pay rate type is daily, then the daily pay, and if pay rate type is hourly, then the hourly pay rate. If the pay rate is zero the employee is typically in a leave status that does not receive pay; for example military leave.
    - TITLE_DESCRIPTION - Title for employee; there are over 500 unique titles.
    - HOME_ORGANIZATION - The home organization code identifies where the employee is primarily stationed. Some employees are "per diem", meaning they work on a daily basis at a variety of locations; other employees may work at multiple locations, but are attributed to one location.
    - HOME_ORGANIZATION_DESCRIPTION - The home organization description identifies the home location by name, typically the name of the school or office.
    - ORGANIZATION_LEVEL - Identifies the type of location the employee works at; for example, administrative office, garage, elementary school, etc.
    - TYPE_OF_REPRESENTATION - The union representation for the employee. NON REP means the employee is not represented by any union.
    - GENDER - Gender of employee.
    - RUN_DATE - The date the data was extracted from SDP data systems.

If you know what each field name stands for, you can work with it without making wrong assumptions.

A more complex record layout or codebook, like the one from the National Inventory of Dams (NID), which contains records on dams in all 50 states (kept by the U.S. Army Corps of Engineers) will look something like this summary of it:

Field	Type	Description

NIDID	text	Nat'l Inventory of Dams ID; for saddle dams or dikes, the NIDID is the same as the main dam.
Dam_Name	text	Dam name
Insp_date	date	Inspection date (converted)
City_02	text	The nearest city, taken from the 2002 data
Year_Comp	number	Year in which the original main structure was completed
NPDP_Hazard	text	Potential hazard to the downstream area resulting from dam failure or misoperation. Taken from Stanford's "National Performance of Dams Program".
Latitude	text	Latitude at dam centerline, in decimal degrees

Always look for the record layout or codebook or data dictionary. If it's not in the dataset. Ask for it. Without it, you can make a lot of mistakes.

Scraping

Extracting data from a website or a PDF document, and creating structured data from the result.

Spreadsheet

A table of data and calculations that can be processed with a spreadsheet program such as Microsoft Excel or OpenOffice Calc or Google Spreadsheets.

Structured data

All data has some structure, but 'structured data' refers to data where the structural relation between elements is explicit in the way the data is organized in a data file.

TSV

Tab-separated values (TSV) are a very common form of text file format for sharing tabular data. TSV files can be imported into and exported from spreadsheet software. TSV files are essentially text files and can be viewed by text editors.

XLS(X)

Microsoft Excel's spreadsheet file format. Older versions use .xls files.

Final note: You can get a full glossary of terms and descriptions of file formats at the Open Data Handbook website http://opendatahandbook.org/glossary/en/. The terms included in this Handout have been extracted from that glossary and summarize for brevity. It comes in handy to bookmark the website address.