CRITERIA FOR AN EXCELLENT OPEN DATA PORTAL

Data Science For the Public Good, Social and Decision Analytics Laboratory, Virginia Tech Biocomplexity Institute
Mark Almanza, Madison Arnsbarger, Jessica Flynn, Millicent Grant, Adrienne Rogers, William Sandholtz, Emily Stark, Samantha Tyner

By making public information accessible to the public online via Open Data Portals, government is more efficient and policy measurements improve. This, in turn, boosts the local economy and businesses. However, this is only possible with the continual improvement and self-critique of the Open Data Portal itself by following the following guidelines.

USER INTERFACE

Organization

The physical set up of the structure of the website and database.

Web pages and datasets are organized consistently
- -Nest folder and links by topics
- (e.g. Utilities > Water Meters > Data set)
- -Site hierarchy is clear, minimizing navigation steps
- -Lists are sorted consistently
- (e.g. Alphabetically, relevance, date, etc.)

Users are given appropriate context
- -Descriptive information is included
- -Explicit cues given on each page for context and organization
- -Additional resources given to aid user exploration

Navigation

The interactive elements of the website guiding the user through the database.

Easily accessible from County webpage
Search function should search by metadata and between many data sets if you have one variable in mind (i.e. be able to search for housing prices from the main page)
Be able to browse by file type (i.e. be able to see all maps) within and between categories exclusively
Should be a button to return to the home page
Be able to toggle through categories

Design

The visual elements of the website's design and flow.

Easy to read
- -Text should be displayed cleanly and legibly, hopefully with high contrast to the background

Consistent visual environment
- -Unless meant to highlight a different environment/functionality

Section headers should be easily viewable from the homepage
- -One shouldn't have to go searching for a primary function i.e. buttons, search bars, etc.

Different types of data and different types of visuals should be differentiable at a glance
- -Use of different color or graphic is recommended

The use of any tool presented to the user should be clear
- -A person who's reasonably internet proficient (not necessarily proficient in statistics) should not be uncertain of the use of tools

Color coding tools is encouraged

DATA

Tidiness

The quality and usability of the datasets.

"Each variable forms a column"
"Each observation forms a row"
Common dataset tidiness tips:
- -Columns represent variables not values
- -One column contains only one variable
- -Variables are stored only in columns, not rows

More tidy suggestions:
- -Keep units in column headers
- -Make sure that variables are internally consistent
- -Try to format special data types (dates, currency, etc) in a coherent, consistent manner across datasets

Source: Wickham, Hadley. "Tidy Data." Journal of Statistical Software [Online], 59.10 (2014): 1 - 23. Web. 1 Jun. 2016

SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Accessability

The ability to get to, understand what is in the and share the data.

Availability:

- -Restrictions are removed and data is license free
- -Datasets are clearly labeled and kept up to date
- -Historical data is included and identified

Safety:

- -Personally Identifiable Information (PIIs) are accounted for (e.g. names, Social Security numbers, biometric records, etc.)
- -Unique Identifiers replace PIIs when possible

Sharing:

- -Code used to create website, datasets and charts are made publicly available
- -Public APIs are availble and usable

Documentation

The background information of the dataset.

Metadata:

- -Column/row is clearly defined
- -Each level of categorical variables are clearly explained

Sources:

- -Publishing page highlighting key interest sources
- -Source of each individual dataset included

Data Sets:

- -Dataset summary is included
- -Clear indication of what time frame the data spans
- -Brief explanation of how data was collected

File Type:

- -Is clearly indicated (csv, xlsx, etc.)
- -Raw data is available
  - -original, unaltered file
  - -sites averages, summaries, etc.
  - -missing values are left blank

Visualization

The visual elements of the dataset.

All Visualizations:

- -Are clearly labeled throughout (titles, axes, legends, etc.) so there is no confusion about what information is being shown.
- -Remain true to the data. They do not distort the information presented.
- -Avoid “chartjunk:” visual elements that do not add to wor may even distract from the viewer’s comprehension of the information presented.
- -Optional: Contribute to the overall understanding and utility of the data source they show. Tell a story about the data.

Creating Visualizations:

- -Users can access the data visualization creation tools quickly and easily.
- -Tools provided for visualization are clearly labeled and easy to navigate.
- -Optional: Data visualization options presented to the user are matched appropriately to data type. (e.g. box plot is not an option for categorical data, but bar charts are)

Source: Tufte, Edward. The Visual Display of Quantitative Information, 2nd Ed, Graphics Press (2001)

Citations
[1] Open Data Portal Requirements. (n.d.). Retrieved June 03, 2016, from http://labs.centerforgov.org/open-data-portal-requirements/
[2] Open Data Policy Guidelines. (n.d.). Retrieved June 03, 2016, from http://sunlightfoundation.com/opendataguidelines/
[3] City Open Data Policies. (n.d.). Retrieved June 3, 2016, from http://www.nlc.org/Documents/Find City Solutions/City-Solutions-and-Applied-Research/CSAR Open Data Report FINAL.pdf
[4] Lynch, P. and Horton, S. (2011) Web Style Guide Retrieved from http://webstyleguide.com/wsg3/4-interface-design/3-interface-deisgn.html

SDAL SOCIAL & DECISION ANALYTICS LABORATORY