SDAL — SOCIAL & DECISION ANALYTICS LABORATORY

Data Science for the Public Good

CRITERIA FOR AN EXCELLENT OPEN DATA PORTAL

Mark Almanza (Virginia Tech), Madison Arnsbarger (Virginia Tech), Jessica Flynn (Cornell), Millicent Grant (Iowa State University), Adrienne Rogers (Virginia Tech), William Sandholtz (UC Berkeley), Emily Stark (Austin Peay), Samantha Tyner (Iowa State U) with Gizem Korkmaz, Katie Ziemer, & Stephanie Shipp (SDAL), and Kristanne Littlefield, Arlington Chief Digital Officer

Goals

- "Open Data," which is any publicly available information, has the potential to improve government efficiency, advance local businesses, and facilitate more frequent and deeper use of government data.
- Arlington County asked students in the Data Science for the Public Good Program (DSPG) at SDAL for feedback regarding their newly-released Open Data Portal.
- The students developed a set of criteria that could be used to develop and improve any open data portal to aid in the evolution and improvement of general Open Data Portal standards.

Criteria

USER INTERFACE

Organization
The physical set up of the structure of the website.

- Web pages and datasets are organized consistently and logically
  - File hierarchy is clear and lists are sorted properly
- Users are given appropriate context
  - Descriptive information included

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

- Easily accessible from the main webpage
- Explicitly labelled button to return to Open Data Homepage
- Ability to search and move through files based on file type, category, or keyword

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

- Easy to read
- Consistent visual environment
- Section headers should be easily viewable from the homepage
- Different types of data and different types of visuals should be differentiable at a glance
- The use of any tool presented to the user should be clear
- 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 in only columns, not rows;
More tidy suggestions:
- Keep units in column headers; make sure variables are internally consistent; format special data types in a consistent manner.

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

- Restrictions/licenses removed
- Headers give full explanation of dataset
- Personally Identifiable Information (PIIs) are accounted for
- Public APIs and code are available and usable

Documentation
The background information of the dataset

- Dataset contents are described in clearly labeled metadata
- All sources are clearly published
- Dataset summaries are available
- Download File Type is clearly labeled

Visualization
The visual elements of the dataset

All Visualizations:
- Clearly labeled
- Remain true to the data
- Avoid "chartjunk"
Creating Visualizations
- Users can access the data visualization creation tools quickly and easily
- Tools provided for visualization are clearly labeled and easy to navigate

Website Design

- A simple palette of high contrast colors should make up the bulk of the site, leaving many colors available as accents.
- The site's color palette should remain consistent throughout the site to aid in navigation and to appear more unified.
- An Open Data Portal should be easy to read, with text in a font and color that are easy to look at for extended periods of time.

Aa

Data Tidiness

Name	Sex	Date of Birth	Height (in)	Start Year	End Year
Amanda	F	03/07/1968	66	1990	1996
Bradley	M	11/05/1971	70	1994	1999
Connor	M	07/17/1964	68	1988	1992
Danielle	F	10/22/1970	67	1997	2005

- This table is in a tidy format: each row represents a different observation, and the columns record various characteristics about each person such as date of birth, sex, or height.
- If a column contains a range of values, the column should be split into two separate columns - one for the minimum, and one for the maximum values of that range.
- While it is open for debate whether the responsibility of tidying data is that of the Open Data Portal administrator, depending on the goal and maturity of the organization supporting the portal, making the data tidy would facilitate better and more frequent use of the content.
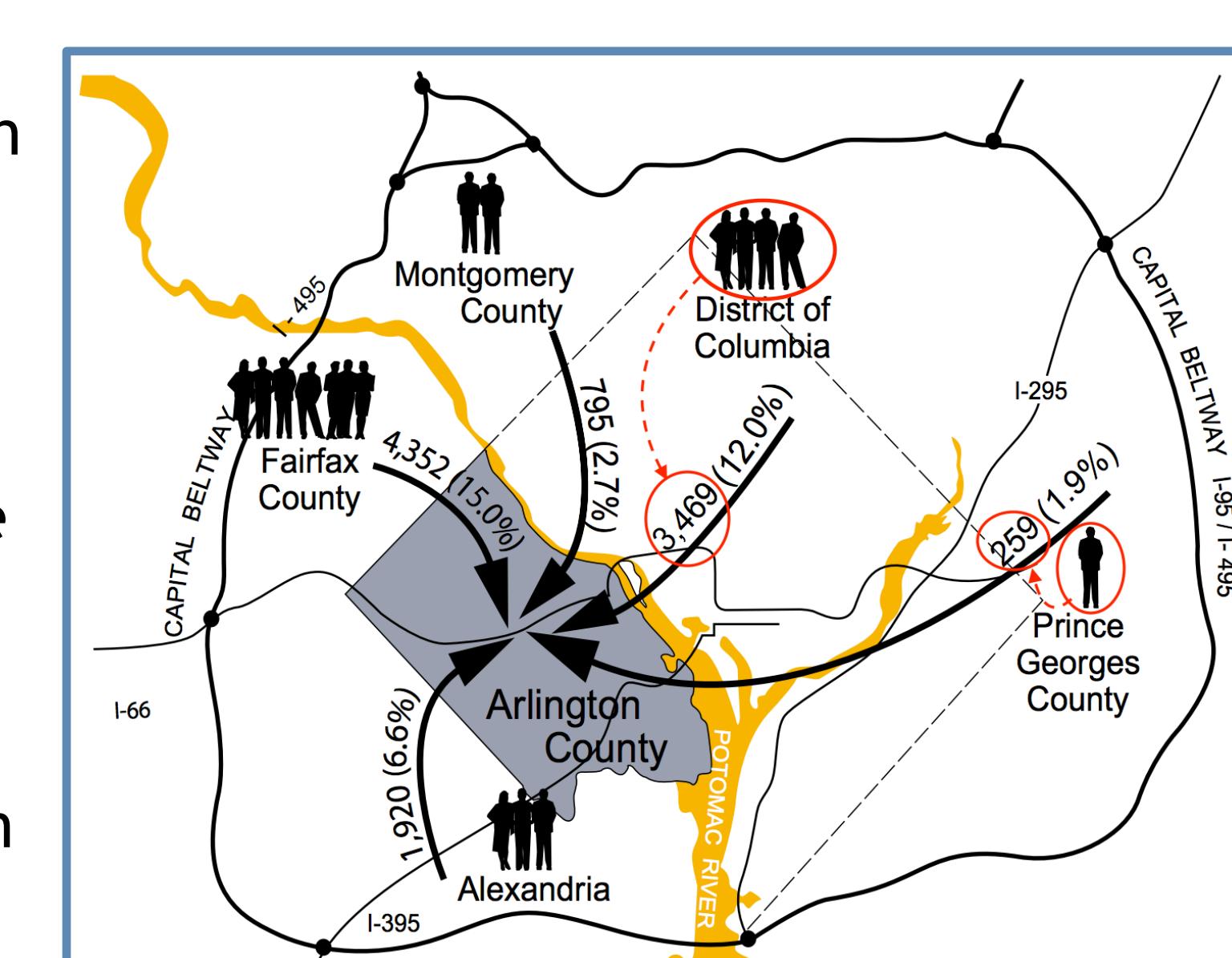
Data Documentation

Publishing Departments: 311, Airport (SFO), Arts Commission, Assessor-Recorder, Board of Supervisors, Building Inspection, Citizen Complaints, City College, Community Investment and Infrastructure, Controller, Economic & Workforce Development, Elections, Environment, Ethics Commission, Film Commission, Fire Department, GSA - City Administrator's Office, Health Service System, Human Resources, Human Rights Commission, Human Services Agency, Mayor, Municipal Transportation Agency, Planning, Police Department, Port, Public Health, Public Utilities Commission, Public Works, Recreation and Parks, Rent Arbitration Board, Technology, Treasurer-Tax Collector, Other

- Proper documentation answers questions about where the data came from and how it was collected.
- Displaying a list of commissions, agencies, and other government organizations that have shared their data on the site may enhance the credibility of the portal.
- The description of each dataset should contain information about the origin of that data.

Data Visualization

- The "junk" items here are the people icons, which represent the flow of people migrating into Arlington County from surrounding areas in 2012.
- Although the number of people coming from DC is more than 13 times the amount of people coming from Prince Georges County, there are only 4 person icons to represent that difference. This misleads the viewer.
- Better options would be to remove the people and make the numbers more prominent, or to have a legend indicating how many people each person icon represents.

References

[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-design.html

VirginiaTech Biocomplexity Institute

bi.vt.edu/sdal