Data Visualisation for Analysis in Scholarly Research

British Library Digital Scholarship Training Programme
September 2015
Mia Ridge @mia_out

While we’re getting started...

• Check that the mouse on your laptop works and that you can get online with the browsers Firefox or Chrome
• Unzip (‘extract’) the file containing the slides and exercise handouts and copy the folder to your desktop
• Dig out your GMail/Google login details (if you have an account)

Timetable

• 10am Start
• 11:30-11:45 Break
• 13:00-14:00 Lunch
• 15:00 Conclude

Links, sources and further reading:
http://www.miaridge.com/2013/01
Overview

- Introductions; what is data visualisation?
- History and types of visualisations
- Critiquing visualisations
- Visualisations for scholarly analysis
- Dealing with library and historical data
- Planning and designing visualisations

Data visualisation is the graphical display of quantitative or qualitative information to create insights by highlighting patterns, trends, variations and anomalies.

From this...
...to this

About me

Tool from http://neatline.org/

Introductions

• In a sentence or two, what’s your interest in data visualisation?
  – What kinds of data do you work with?
  – What’s the goal of any visualisations you’re interested in creating?
  – Do you have any potential users in mind?
What is data visualisation?

'sense-making (also called data analysis) and communication' (Stephen Few)

'...showing quantitative and qualitative information so that a viewer can see patterns, trends, or anomalies, constancy or variation' (Michael Friendly)

'...interactive, visual representations of abstract data to amplify cognition' (Card et al)

Visualisations as intersection of format and purpose

• Product or process? Exploratory or explanatory: find new insights, or tell a story?
• Static or interactive; print or digital?
• Pragmatic, emotive?
• 'Distant reading' - focus on the shape rather than detail of a collection

Data visualisation can help you...

Explore your data

Explain your results
Exploring data

HISTORY AND TYPES OF VISUALISATIONS

Joseph Priestley, 1769

A NEW CHART OF HISTORY
John Snow’s cholera map, 1854

Florence Nightingale’s petal charts, 1857

Charles Minard’s figurative map, 1869

The old tube map

Harry Beck, 1931

Web 2.0 and the mashup, 2006
Exploring words

http://www.jasondavies.com/wordtree/

Visualising images and video, 2012

http://www.flickr.com/photos/culturevis/5883371358/  
Data types

- quantitative
- qualitative
- geographic
- time series
- media
- entities (people, places, events, concepts, things)

CRITIQUING VISUALISATIONS

'sentiment'

Olympics in review – the big Twitter sentiment timeline
A sample of publication printing locations 1534-1831 (British Library data) http://bit.ly/W9VM7D

Network visualisations
http://fredbenenson.com/blog/2012/12/05/the-data-behind-my-ideal-bookshelf/

Exercise 1: network visualisations
Instructions on the hand-out.
N-grams

Exercise 2: comparing N-gram tools

Topic modelling
Other forms of text analysis

Entity recognition:
turning text into things

Exercise 3: trying entity recognition

Instructions on the hand-out.

Entity recognition examples

Named Entity Recognition:

1. Speaking at a UN conference in Sendai, Japan, on Monday, said 90% of buildings in Port Vila had been
damaged or destroyed by the category five storm, which saw winds of up to 255km/h (159mph). DOROTHY

2. "It's a well back for the government and for the people of Vanuatu, he said.

3. "After all the development that has taken place, all this development has been wiped out." DOROTHY

4. Communications in the Port Vila province of the island have now been 'almost fully restored', according

5. to telecommunications provider Digicel, allowing information to flow more freely to and from disaster

6. Several countries have also pledged additional aid and funding for the stricken island nation.

7. The Australian foreign affairs minister, Julie Bishop, pledged A$7m in support, and New Zealand has

offered $2.5m.
Scholarly data visualisations

- Visualisations as 'distant reading' where distance is 'a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection' (Moretti, 2005)
- Inspiring curiosity and research questions
- But - which questions do they privilege and what do they leave out?

Exercise 4: explore scholarly visualisations

Pair up and discuss together before reporting back.

Instructions on the hand-out.
Visualizing Emancipation

http://www.americanpast.org/emancipation/

Mapping the Republic of Letters

http://www.stanford.edu/group/toolingup/rplviz/rplviz.swf

GAPVis

http://gap.alexandriarchive.org/gapvis/index.html
Digital Harlem

http://digitalharlem.org

Digital Public Library of America

http://dp.la/

Orbis

http://orbis.stanford.edu/#mapping
Lost Change

http://tracemedia.co.uk/lostchange/

State of the Union

Read Obama's State of the Union in context! Made by Ben Schmidt Using Bookworm

The State of the Union in Context

This site lets you use a Bookworm database to read Obama's State of the Union in the context of all the other State of the Union messages given by American presidents. For any word in the message, just click on the text; the word will turn red. If you want to search a two-word phrase, just highlight both words. (You can search for phrases longer than two words.)

Mr. Speaker: Mr. Vice President, Members of Congress, my fellow Americans

We are richer than ever in this new century. Progress that started with some tough choices in our past—choices that included a two-generation fighting two wars and wars were that we balanced our budget, and cut taxes, and invested in research and development—has finally paid off.

Now, after a tremendous year, for America, our economy is growing and underlying index at the breakneck pace since 1995. Our unemployment rate is now the lowest since 1969. Our unemployment rate is now the lowest since 1969. Moreover, we are now home to the Gulf of Mexico, as we always did in a short 30 arts.

Comments or questions?

http://benschmidt.org/poli/2015-SOTU
ISSUES WITH HISTORICAL, CULTURAL DATA

Considerations for historical data

Commercial tools often assume complete, born-digital datasets – no missing fields or changes in data entry over time

- Historical records often contain uncertainty and fuzziness (e.g. date ranges, multiple values, uncertain or unavailable information)
- Includes metadata, data, digital surrogates

Messiness in historical data

- ‘Begun in Kiryu, Japan, finished in France’
- ‘Bali? Java? Mexico?’
- Variations on USA:
  - U.S.
  - U.S.A
  - USA
  - United States of America
  - USA?
  - United States (case)
- Inconsistency in uncertainty
  - U.S.A. or England
  - U.S.A./England?
  - England & U.S.A.
When were objects collected?

http://ibm.co/OS3HBla

Computers don't cope

Preparing data for visualisations

Historical data often needs manual cleaning to:

- remove rows where vital information is missing
- tidy inconsistencies in term lists or spelling
- convert words to numbers (e.g. dates)
- remove hard returns and non-ASCII characters (or change data format)
- split multiple values in one field into other columns (e.g. author name, date in single field)
- expand coded values (e.g. countries, language)
Data Preparation

• Generally needs to be in tables, one row per item, one column per value

• Might need to calculate values in advance

• Data should be made as consistent as possible with tools like
  – Excel
  – OpenRefine [http://openrefine.org](http://openrefine.org)

Open Refine

…but be careful
Purpose, data, audiences (revision)

- Intersections of format and purpose
- Data types: quantitative, qualitative, geographic, time series, media, entities (people, places, events, concepts, things)
- Static, interactive; print, digital; product, process
- Exploratory, explanatory: find new insights, or tell a story? Pragmatic, emotive?
Choosing a structure

See relationships among data points

- Scatterplot
- Matrix
- Network diagram
Compare a set of values

- Bar chart
- Bubble chart
- Histogram

Track change over time

- Line graph
- Stack graph

See the parts of a whole

- Pie chart
- Treemap
Exercise 5: create a chart using Google Fusion Tables

- Instructions on the hand-out

- If you would rather try an exercise in Excel, see instructions for creating simple graphs with Excel's Pivot Tables and Tate's artist data

DESIGNING VISUALISATIONS

Worst practice in data visualisations

Source: http://www.forbes.com/sites/naomirobbins/2013/01/03/deceptive-donut-chart/
Worst practice in data visualisations

Source: https://twitter.com/altonnc/status/293392615225823232

Best practice for design

• How effectively does the visualisation support cognitive tasks?
• The most important and frequent visual queries/pattern finding should be supported with the most visually distinct objects

Visually distinct objects

• Colour (hue, lightness)
• Elementary shape (orientation, size, elongation)
• Motion
• Spatial grouping
Dealing with complex data

- Find a visualisation type that can harbour the data in a meaningful way or reduce the data in a meaningful way.
  - e.g. go from individual values to distribution of values
  - e.g. introduce interaction: overview, zoom and filter, details on demand (Ben Shneiderman)
Do you really need a visualisation?

- Use tables when:
  - the document will be used to look up individual values
  - to compare individual values
  - precise values are required
  - the quantitative info to be communicated involves more than one unit of measure

- Use graphs when:
  - the message is contained in the shape of the values
  - the document will be used to reveal relationships among values

Publishing visualisations

- How can you contextualise, explain any limitations of your visualisations? e.g.
  - provenance and qualities of original dataset;
  - what you needed to do to get it into software (how transformed, how cleaned);
  - what’s left out of the visualisation, and why?

Tools that don't require programming

- Excel
- Google Fusion Tables, Google Drive
- IBM Many Eyes
- Tableau Public
Exercise 6: geocoding data and creating a map using Google Fusion Tables
• Instructions on the hand-out

Review: planning a visualisation
• With a dataset in mind, consider...
• Exploratory or explanatory? Static or dynamic? Small- or large-scale?
• Choose a type of visualisation (map, timeline, chart, etc)
  – Is your dataset in a suitable format for your visualisation type? How can you clean it?
  – Is more cleaning or transformation needed? You may need to iterate with different versions of your data

If all else fails...
• Sketch out your visualisation on paper to test it
• Iteration is key, and...
• Stubbornness is a virtue!
Exercise 7: taking things further

• Try more visualisations
• Sketch visualisation ideas
• Try visualisation tools
• Instructions on the hand-out

Questions or comments?

References and finding out more


Thank you!