(data) Visualization

for the digital humanities and beyond

Clement Levallois

v.1: January 2012 v.2 (latest): April 2013

(data) Visualization

for the digital humanities and beyond

Why visualizations?

"The ability to collect, store, and manage data is increasing quickly, but our ability to understand it remains constant." (**Ben Fry** 2005). In the humanities, this translates in:

- Expanding databases of texts, pictures, movies, sounds, webpages, files from various software
- Growing collections of artifacts in museums and archives in libraries.

The volume and complexity of the data produced and stored by the digital humanities means that they remain largely untapped. How to start exploring them? **John Tukey**'s proposal for **"Exploratory Data Analysis"** (**EAD**), which he developed in the course of the 1960s onwards, proves to be useful.

This approach suggests that after a suitable preprocessing of the data, the naked eye is a powerful instrument to generate insights from very large datasets. Tukey: "[T]he picture-examining eye is the best finder we have of the wholly unanticipated." (1980).

As the Wikipedia entry for "Exploratory Data Analysis" develops, this is an approach to "analyze data sets to summarize their main characteristics in easy-to-understand form, often with **visual** graphs, without using a statistical model or having formulated a hypothesis". Developments in the last decade provide powerful tools to generate insightful visualizations. **We present some of these developments in the following pages**.

Fry and Tukey

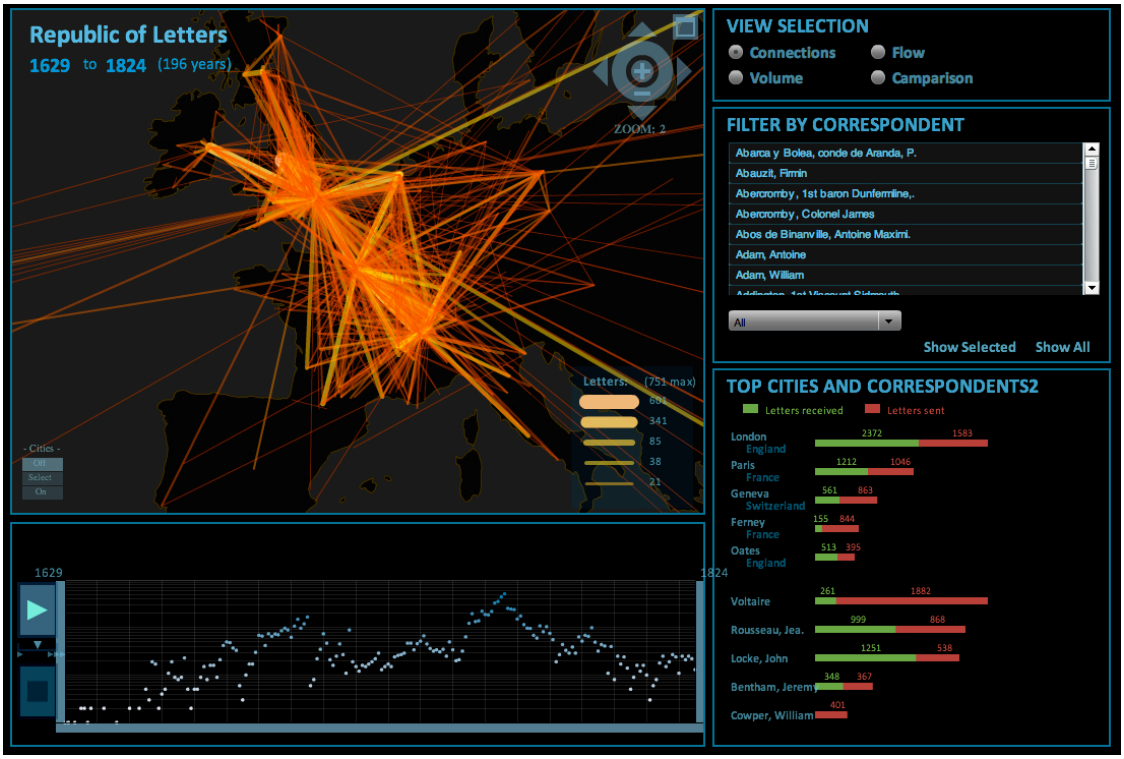
Ben Fry (1975-) graduated in 2005 from the Aesthetics + Computation Group at the MIT Media Laboratory. His PhD thesis provides a theoretical and operational framework for the visual exploration of data. With Casey Reas, Fry developed "Processing", a Java-based language very popular with designers and artists for visual, data-based, interactive projects.

John Tukey (1915-2000) was a leading statistician of the twentieth century best known today for his fast Fourier transform (FFT) algorithm and his invention of the "box-plot" graphing technique.

What visualizations?

3 projects illustrate the diversity of purposes and meanings which visualizations can serve in the digital humanities.

1. **"Mapping the Republic of Letters" (Stanford Humanities Center, ongoing)**

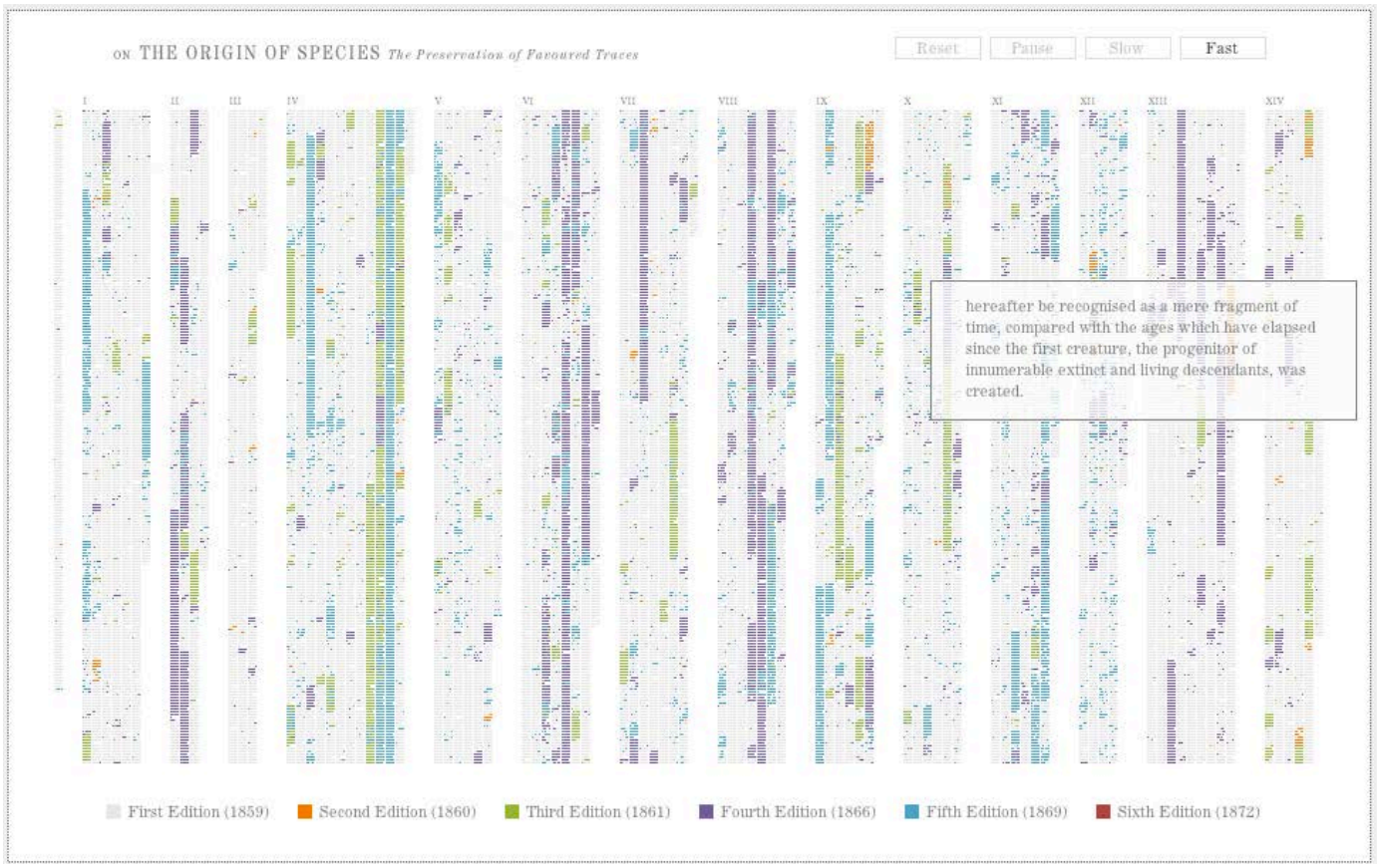


This project demonstrates how heterogeneous data sources in large volumes can be synthesized in an interactive display:

"The collection of metadata grew quickly to include travel records, library catalogs, and the circulation of scientific instruments as well as correspondence. The number of faculty and graduate students contributing data and case studies to the project has steadily grown as well, creating a humanities laboratory where graduate students, undergraduates, and faculty, humanists and technical experts regularly share the results of their work and discuss common problems in the visualization of humanistic data."

The visualization is developed in Flash.
source: https://republicofletters.stanford.edu/

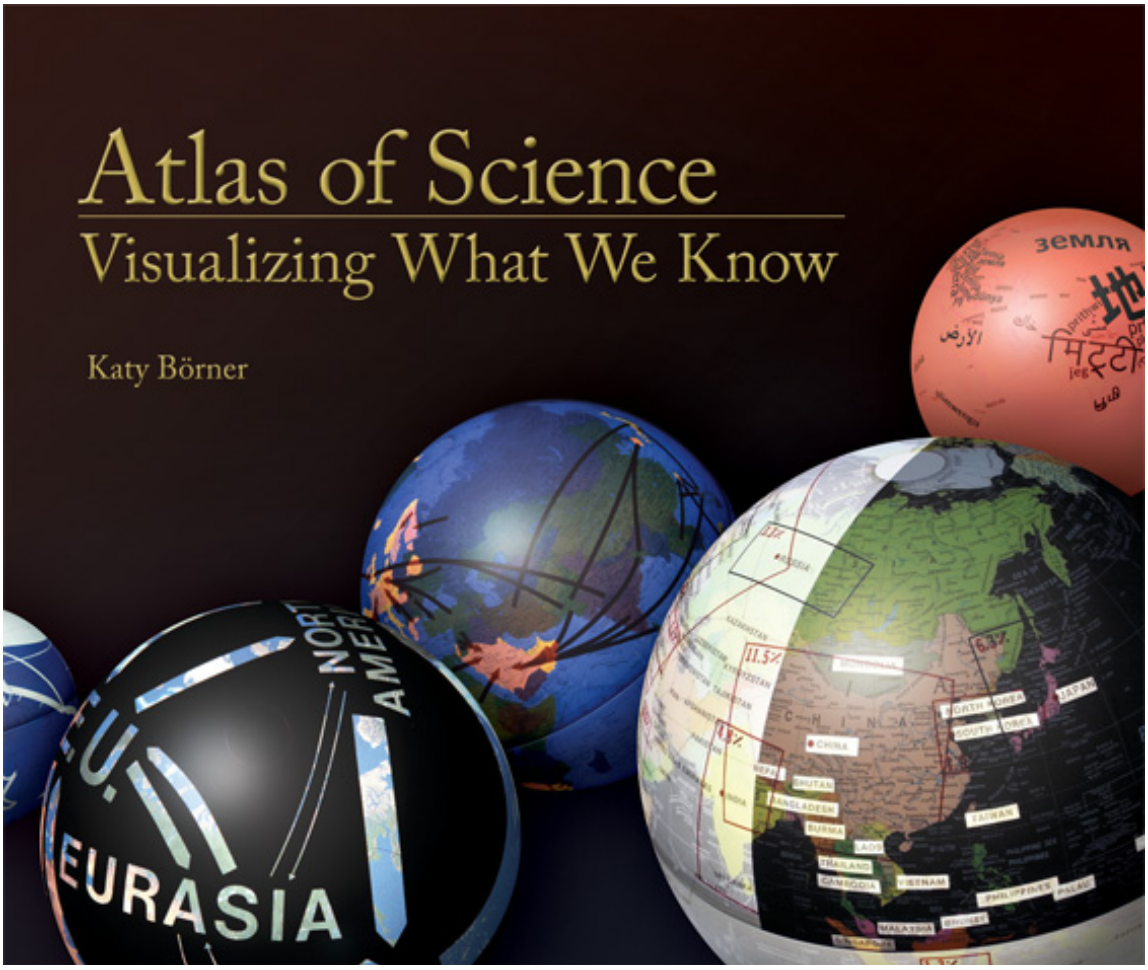
2. "On the Origin of Species: The Preservation of Favoured Traces" (Ben Fry, 2009)



This snapshot of the interactive project shows how an aggregate view can readily show the extent and patterns of variations between editions, while the access to micro details is preserved thanks to a fish-eye effect revealing individual sentences from the text. This work is also an excellent example of the valorization of a dataset which took an immense effort to constitute, and is still is costly to maintain (http://darwin-online.org.uk/)

The visualization is developed with Processing.
source: http://benfry.com/traces/

3. Atlas of Science (Katy Börner, 2010)



This book sums up a series of exhibitions on "Places & Spaces: Mapping Science" featuring the diversity of maps used to represent what we know. The result is a large tome in full colors, making the case that the community of map makers has produced a large body of techniques and practices – which can now be reused for diverse research projects. The form of the output – book and exhibitions – also testifies that digital humanities are not "screen bound" and that many forms of engagement with the audience can be pursued.
source: http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=12344

Visualizations: how?

1. Very small sample of click and point software / web platforms

VosViewer (Leiden University)
Takes network data or textual data as input, generates maps of word clusters and heatmaps.

Gephi (Gephi Consortium)
Takes network data as input, generates network maps "in real time". Dynamic networks and a large number of user defined layouts and plugins are supported.

NodeXL (Social Media Research Foundation)
Functions as an Excel spreadsheet. Takes network data as input, generates maps as output.

Tableau Public (Inc)
Takes a variety of data as input, generates charts and geographical maps as output. Online platform, data must be public and can be shared.

ManyEyes (Microsoft)
Microsoft's online platform. Accepts many data formats, returns a wide variety of outputs (maps, charts, wordles, etc.)

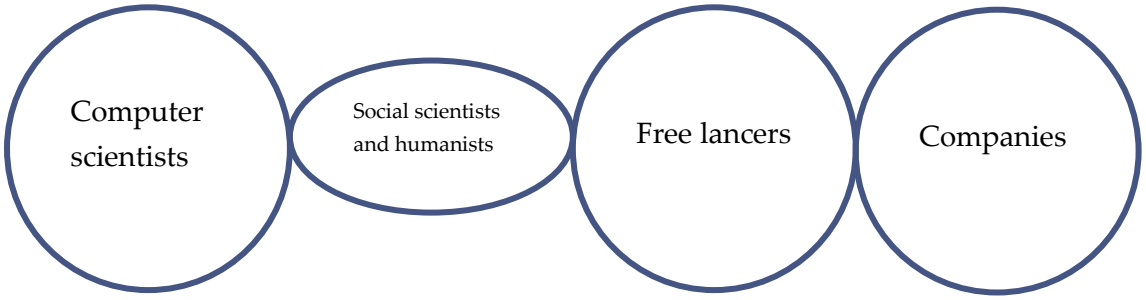
Google Fusion Tables
Takes geolocalized data as input, generates Google maps with layers of information as output.

2. Programming languages

Programming takes time to learn, but the benefits are substantial with the full control gained over the transformations applied to the data – which is lacking in click-and-point software. Note: some software listed above (like Gephi or NodeXL) are open source: this means that one can use them in their "click and point" version, or look inside and modify their code as needed.

For visualizations on the web (pages in an Internet browser), it exists a variety of Javascript libraries to create stunning things (a "library" is simply a collection of pre-written code that gives you a big head start, you still have to add some code you write yourself to get to the result). There are dozens or hundreds of javascript libraries for data visualization, and libraries written in other programming languages too (Processing, R, Java, python, etc.). A neat selection of these tools can be found here: http://selection.datavisualization.ch/

Visualizations: by whom?



Visualizing a dataset entails many different forms of expertise, though a number of software (like VosViewer) and programming languages lower the threshold for single users to conduct an analysis from start (data collection) to finish (sharing insights drawn from the visualization). The field is a composite of scientists from different horizons, **however the social scientists and researchers in the humanities are very much underrepresented**. A subjective, very restricted list provides a starting point:

Academics
[Too numerous to cite. I mention just 3 "hubs" with particularly diverse connections in the community]
Enrico Bertini (@filwd), NY Polytech
Elijah Meeks, Stanford
Katy Borner, Univ. Indiana

Companies
A list is actively maintained on www.quora.com, under the question: "Who are today's leading data visualization companies?"

Newspapers
The NYT and the Guardian are leaders in data visualization:
http://nytlabs.com/
http://www.guardian.co.uk/news/datablog

Professionals (twitter handle)
Jerome Cukier (@jcukier)
Noah Iliinsky (@noahi)
Andy Kirk (@visualizingdata)
Santiago Ortiz (@moebio)
Kim Rees (@krees)
Moritz Stefaner (@moritz_stefaner)
Jer Thorp (@blprnt)
Jan Willem Tulp (@janwillemtulp)
Nathan Yau (@flowingdata)

Specialists in critical reviews of data visualization projects
VizWiz: http://vizwiz.blogspot.com/
the Why Axis: http://thewhyaxis.info/
Junk Charts: http://junkcharts.typepad.com/
Datavisualization.ch

Where? Events and references on visualization

Events by / with computing engineers
IEEE Visualization (VisWeek)
IEEE Information Visualization Conference (InfoVis)
Eurographics/IEEE Symposium on Visualization (VisSym/EuroVis)

Events by / with graphic designers, artists, businesses
EyeO Festival (US)
Malofiej (Europe)
SEE (annually in Wiesbaden, Germany)

Twitter
Professionals in data visualization are very active news sharers (see their twitter accounts on the previous page). Hashtags to follow: #dataviz, #datavis, #infovis

Selected bibliographical references

Börner, K. (2010). Atlas of Science: Visualizing What We Know. MIT Press, Cambridge Mass.
Fry, B. (2005). Computational Information Design (PhD diss.). MIT Press, Cambridge Mass.
Fry, B. (2008). Visualizing Data: Exploring and Explaining Data with the Processing Environment. O'Reilly Media.
Steele, J., & Iliinsky, N. (Eds.). (2010). Beautiful Visualization: Looking at Data through the Eyes of Experts. O'Reilly Media.
Tufte, E. R. (1997). Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press.
Tukey, J. W. (1980). We need both exploratory and confirmatory. The American Statistician, 34(1), 23-25.