Crash Course Digital Humanities 2014

Day 3 - Preparing Data

Fernie Maas, Jan Hein Hoogstad, Marijn Koolen

22 October 2014

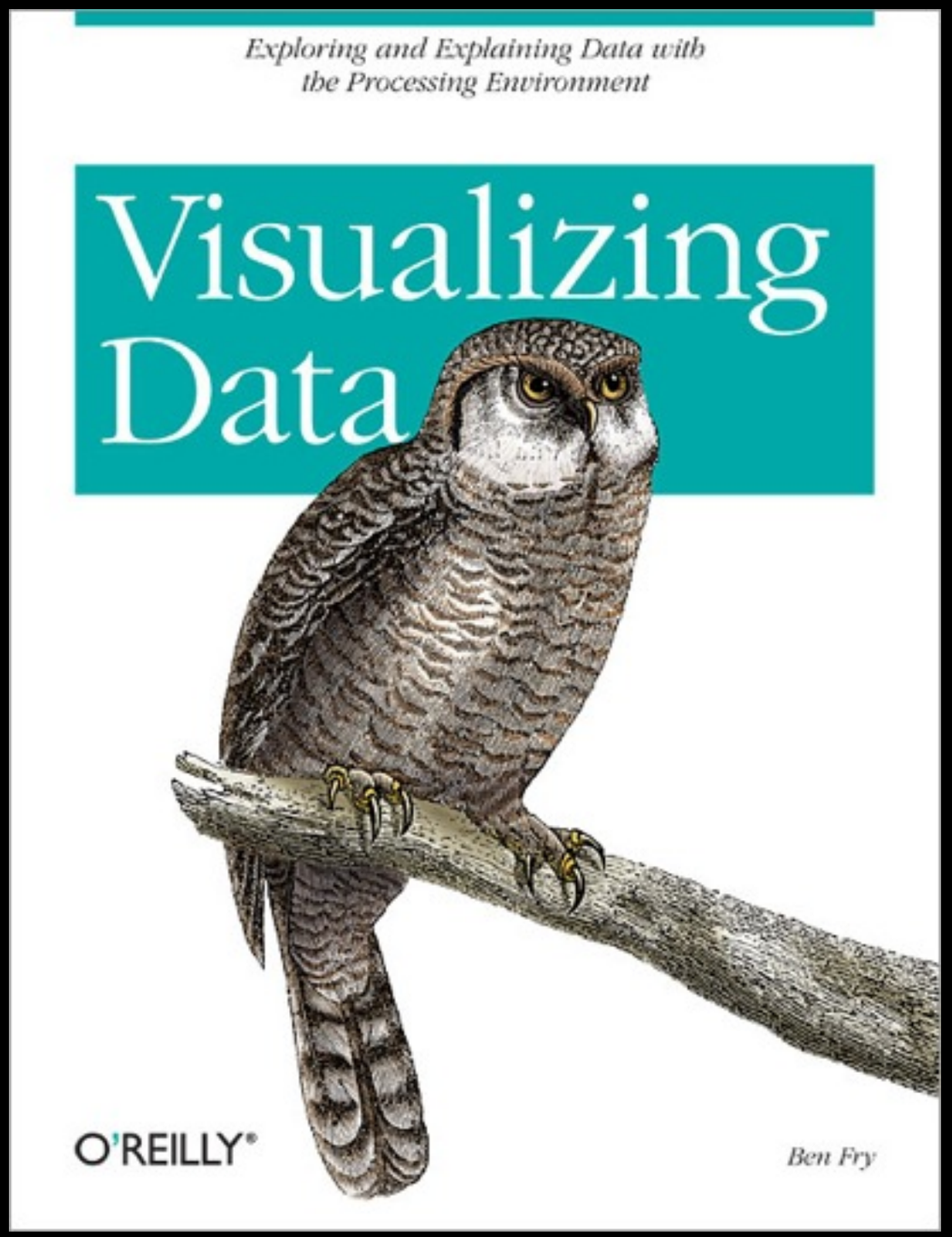
Waag, Amsterdam, Netherlands

Today

- 13:00-15:00 Jelle Zuidema - Regular Expressions
- 15:00-15:30 Coffee break
- 15:30-16:30 APIs and NY Times data set
- 16:30-17:00 Discussion

Working With Data

- Visualizing Data - Ben Fry (2008)
  - A framework for understanding data
  - A process of 7 steps



Seven Stages

- Process of understanding data:

1. Acquire: how to get data from sources
2. Parse: identify and label individual bits of data (encoding)
3. Filter: remove or extract data that match specific criteria
4. Mine: discern patterns, statistics
5. Represent: choose visual model (bar graph, tree, ...)
6. Refine: improve basic representation (colours, zoom)
7. Interact: add methods for manipulation (control visibility)

Command Line Tools

- Computational primitives
- Related to scholarly primitives identified by John Unsworth (paper is on the website)
- Command line: breaking up scholarly work in small steps

Part I - Jelle Zuidema

Step 1: Acquiring Data

- Where to find data?
  - Many sources online, offering range of access methods
  - Typical: search & browse
  - We focus on APIs
    - Application Programmer Interface
    - Web standard for programmatic database access
    - Extract data using REST queries (more in a minute)

Which API?

- There are hundreds of thousands of APIs on the web
- National governments, archives, libraries, museums, social network sites, companies
- available data differs across providers
- Examples: NY Times, Echonest, Europeana, Rijksmuseum, Facebook, Twitter, Open Library, KB, Marvel Comics, …

API Requests

- APIs allow you to send a query and get results back
- queries have standard format
  - API_url?query_parameters
  - Example with Europeana API: query “bribery”:
  - http://europeana.eu/api/v2/search.json?wskey=bDyxirp5R&query=bribery&start=1&rows=100&profile=standard

API Response Format

- In browsers it is usually HTML (for display)
- Can be in other formats
  - some APIs allow you to specify format
  - XML: like HTML but more flexible, machine-to-machine data exchange
  - JSON: simple format, becoming worldwide standard

Clients

- Client is often, but not necessarily, a browser
  - can be any program
  - We can use command line to acquire data through APIs

Access To APIs

- How do you get access to APIs?
- Which online sources have data that is relevant to your research?
- Which of those sources provide APIs?

Why APIs

- Why would we want to use APIs instead of standard search and browse interfaces?

Command Line Tools

- Tutorials
  - Command line crash course
  - Unix for poets
- Tools
  - Sublime Text 2 is a great text editor for working with scripts (e.g. syntax highlighting)

Wrap Up

- Thinking in steps - read Unsworth (on website)
- Commands often trip you up (that's good)
- Scripts
  - document and explicate process,
  - structure thinking,
  - allow automation, iteration

Tomorrow

- Case Study 2: Riddle of Literary Quality
- Speaker: Andreas van Cranenburgh
- Automatic genre identification in novels
- Tools: Anaconda (& Python programming language)