DOI: 10.5281/zenodo.1443310

What is data in the Humanities?

Creation, discovery, and analysis

Daniel Paul O’Donnell
University of Lethbridge

Traditionally, humanists resist speaking of data

- “Primary sources” = Texts, artifacts, objects of study
- “Secondary sources” = Works of other scholars
- “Readings” (1) = Passages, extracts, quotations for interpretation or support
- “Readings” (2) = Interpretation, the end product of research (literary study)

Traditionally, humanists resist speaking of data

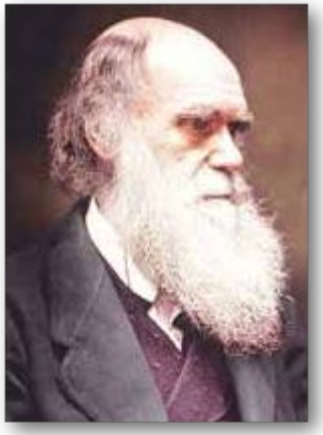
- Our definitions are highly contingent
  - “Primary source” in one context, can be the “secondary source” in another (and vice versa)
  - Or simultaneously “Primary” and “Secondary” (e.g. a critical edition)
- Also hard to constrain

  “a historical text, simultaneously primary and secondary. As Christine Borgman notes, “[a]lmost any document, physical artifact, or record or human activity can be used to study culture” and arguments proposing previously unrecognised sources (“high school yearbooks, cookbooks, or wear patterns in the floors of public places”) are valued acts of scholarship”

(Borgman 2007)

How does data work in other fields?

- Resistance makes sense, because Humanities data is different from other forms of data
- In other domains, “data” (“given things”) is more properly “capta” (“taken”): generated through experiment, observation, and measurement
- Think about Darwin and his work in the Galapagos Islands
    - What is his data?



How does data work in other fields?

- Resistance makes sense, because Humanities data is different from other forms of data
- In other domains, “data” (“given things”) is more properly “capta” (“taken”): generated through experiment, observation, and measurement
- Think about Darwin and his work in the Galapagos Islands
    - What is his data?

The finches?

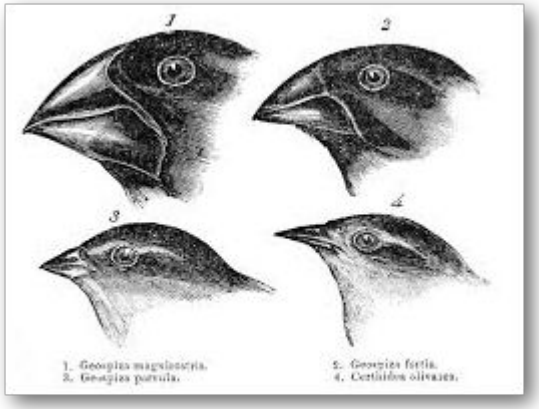
How does data work in other fields?

- Resistance makes sense, because Humanities data is different from other forms of data
- In other domains, “data” (“given things”) is more properly “capta” (“taken”): generated through experiment, observation, and measurement
- Think about Darwin and his work in the Galapagos Islands
    - What is his data?

The notes about the finches?

How does data work in other fields?

- In fact, in the sciences, it is the notes.
- “Data” = “represent[ation of] information in a formalized manner suitable for communication, interpretation, or processing” (NASA 2012); “the facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors” (NRC 1999)



The notes about the finches.

In Humanities, “Data” is arguably mostly “Finch”

- In other humanities, “data” is both “data” and “capta” (given and taken), but more often “data”
- No protocols for preserving our notes (and in most cases nobody would be interested in them)
- Often unique and usually provisional, depend on broader understandings of purpose, context, and form that are themselves open to analysis and modification

Mostly individual finches, maybe something about Darwin, maybe something from our notes

In Humanities, “Data” is arguably mostly “Finch”

- Interesting proof: Humanities “data,” unlike science “data” is almost all practically and theoretically non-rivalrous.
- Humanities researchers rarely have an incentive (or capability) to prevent others from accessing their raw material.
- 200 years of Jane Austen studies based on five main pieces of data.

Mostly individual finches, maybe something about Darwin, maybe something from our notes

DH has the potential to bring new approach to data

- We can now have “capta” (intermediate “observations” extracted algorithmically from large data sets that are then require interpretation)
- We can now work across complete historical or geographic corpora: all known nineteenth-century English periodicals; every surviving tract from the U.S. Civil War
- Introduces the possibility of deductive work
- Makes questions such as sample bias more important than when you worked inductively from the collections you could access

Does this invalidate previous work?

- New forms of data introduce new types of techniques and questions:
  - Falsification as standard of proof?
  - Questions of sampling practice and bias
  - Lab books?
  - Requirement to share data protocols?
  - Requirement to share raw data?
  - Hypotheses rather than theses?
  - Report null results?

Does this invalidate previous work?

Ian Watts, The Rise of the Novel (1957)

- Five novels by three novelists (Defoe, Richardson, Fielding)
- All male, all white, all eighteenth century, all English

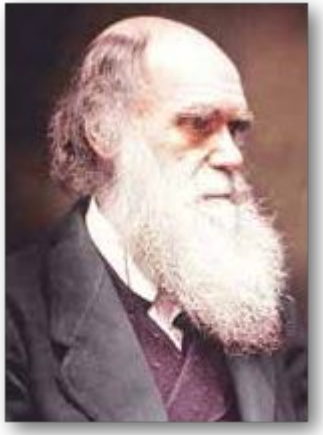
Matt Jockers (2013)

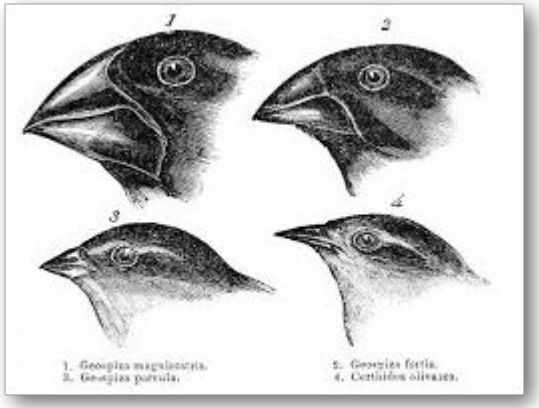
What are we to do with the other three to five thousand works of fiction published in the eighteenth century? [...]Watt had no yardstick against which to make such a measurement. He had only a few hundred texts that he had read. Today things are different. The larger literary record can no longer be ignored: it is here, and much of it is now accessible.

In fact, it means enrichment

- “Capta” and “Data” are different approaches that answer different questions
- But working with Capta will require us to be more careful about our Data
  - Watts’s title *Rise of the Novel* makes a historical claim his actual work doesn’t support: really about how Fielding, Defoe, and Richardson fit into genre
  - Access to 5k novels doesn’t invalidate his arguments; but it does call attention to overreach
  - Can’t imagine that he’d not want access to an even broader collection of work; but I’m not sure his argument would have to be much different.

We now have a greater scope for work





Thank you